Generative Modelling of Quantum Processes via Quantum-Probabilistic Information Geometry

Sahil Patel
Faris Sbahi
Antonio Martinez
Dmitri Saberi
Jae Yoo
Geoffrey Roeder
Guillaume Verdon

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-256
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-256.html

December 1, 2022

Generative Modelling of Quantum Processes via Quantum-Probabilistic Information Geometry

by Sahil Patel

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Umesh Vazirani
Research Advisor

05/19/22

(Date)

* * * * * * *

Professor K. Birgitta Whaley
Second Reader

05/19/22

(Date)

Generative Modelling of Quantum Processes via Quantum-Probabilistic Information Geometry

by

Sahil Patel

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Umesh Vazirani, Chair
Professor K. Birgitta Whaley

Spring 2022

Generative Modelling of Quantum Processes via Quantum-Probabilistic Information Geometry

Abstract

Generative Modelling of Quantum Processes via Quantum-Probabilistic Information Geometry

by

Sahil Patel

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Umesh Vazirani, Chair

Generatively modelling properties of a single quantum system can already be computationally expensive, and often, one wishes to model several different scenarios to see how the dynamical or equilibrium properties of a quantum system evolve as certain parameters, such as time, temperature, or the Hamiltonian, are continuously modified. For such parametric families of tasks, there is often an inherent information geometry for this space of tasks and within each task; one would ideally leverage awareness of such a geometry to guide the optimization of generative models from task to task. Here we explore the use of quantum-probabilistic hybrid representations that combine probabilistic generative models with quantum neural networks, paired with optimization strategies which convert between the geometry of the task space and that of the parameter space of our models, in order to achieve an optimization advantage. We specifically study Riemannian metrics defined on the space of density operators, in particular the Bogoliubov-Kubo-Mori (BKM) metric, which can be well-estimated in an unbiased fashion for our class of quantum-probabilistic models, namely quantum Hamiltonian-based models (QHBMs). We show that natural gradient descent with respect to this construction attains quantum Fisher efficiency of parameter estimation. We further present an alternative first-order formulation of mirror descent that is conducive to improvements in quantum sample complexity. We also derive conditional initialization strategies for simulating time evolution processes and equilibrium states for various values of the problem space parameters. We demonstrate both theoretically and numerically that such techniques may enable accelerated convergence to more optimal solutions of quantum generative modelling tasks.

Contents

Chapter 1

Introduction

Understanding of quantum mechanical systems depends on the interaction between theory and experiment. With the rise of computers in the late 20th century, theory has come to include numerical simulation of physical systems beyond the reach of analytic techniques.

While current state-of-the-art simulations are run on classical computers, quantum computers have been proposed to more efficiently simulate quantum systems [8]. Recently, quantum computers have experimentally surpassed the performance of classical computers on the specialized task of simulating the output of random quantum dynamics [3]. In the coming years it is expected that quantum computers will surpass classical computers on increasingly practical tasks.

Many algorithms have been proposed for the task of quantum simulation. A universal algorithm for quantum simulation on quantum computers was developed as early as 1996 [17], where techniques for both closed and open quantum systems were proposed. However, these proposals often require quantum circuits with depths far beyond the reach of today's Noisy Intermediate-Scale Quantum (NISQ) processors [23]. Variational algorithms offer an alternative approach which can reduce the circuit depth requirements for simulation tasks, making them more amenable to near-term hardware at the cost of requiring classical parameter optimization. In particular, these algorithms take the heuristic perspective that has served machine learning well in recent years by defining a loss function on the samples from a quantum computer, and optimizing the parameters of a quantum circuit to minimize that loss [20, 19].

Early efforts on variational quantum algorithms largely focused on pure state simulation, but many physical systems cannot be described by pure states. Instead, most systems exist at non-zero temperature, or are entangled with systems not accessible to the experimentalist [13]. These systems are described by classical probability distributions over pure quantum states and the required combination of classical and quantum information is typically summarized using a density operator. Recently, quantum Hamiltonian-based models (QHBMs) [29] were introduced as a new variational architecture for simulating density operators by combining probabilistic generative models with quantum neural networks. Specific tasks for these class of models include generating the thermal state for a given Hamiltonian and

inverse temperature as well as learning the modular Hamiltonian of an unknown density operator state through query access to the state.

In this work, we extend the theory of QHBMs by exploring the information geometry of these quantum-probabilistic models. We specifically study the Bogoliubov-Kubo-Mori (BKM) metric defined over the space of density operators and leverage the information geometric structure it induces to construct more sophisticated optimization methods for effectively learning parameterizations of QHBMs. First, we derive metric-aware descent algorithms for QHBMs, including in particular the traditional second-order method of natural gradient descent as well as an equivalent first-order formulation of mirror descent, that yield steepest descent updates towards the minimum of a loss function while maintaining a fixed step size in density operator space. Furthermore, we define priors for learning sequences of quantum states with QHBMs, for instance to simulate equilibrium and time evolution processes, which exploit the geometric locality between adjacent states in the sequence. We demonstrate that such techniques informed by the information geometry of QHBMs can enable accelerated convergence to more optimal solutions of variational learning tasks in quantum simulation.

Chapter 2

Background

2.1 Quantum Hamiltonian-Based Models

Quantum Hamiltonian-based models (QHBMs) were originally proposed in [29] as a new class of quantum machine learning models for generative modelling of density operators. We formally define our density operator space $\mathcal{M}^{(N)}$ to be set of all $N \times N$ density matrices $\hat{\rho}$, where the dimension of the corresponding Hilbert space $\mathcal{H}$ is $\dim \mathcal{H} = N = 2^n$, and that of $\mathcal{M}^{(N)}$ itself is $\dim \mathcal{M}^{(N)} = N^2 - 1$. As density operators describe a classical probabilistic mixture over pure quantum states, QHBMs accordingly parameterize this space through a hybridized representation featuring both quantum and classical correlations. In particular, we may view the quantum-probabilistic structure of QHBMs in terms of two equivalent formulations as follows. Under the mixture representation, a density operator $\hat{\rho}_{\boldsymbol{\Omega}} \in \mathcal{M}^{(N)}$ is parameterized as

$$\hat{\rho}_{\boldsymbol{\Omega}} = \sum_{\boldsymbol{x}} p_{\boldsymbol{\theta}}(\boldsymbol{x}) \hat{U}_{\boldsymbol{\phi}} \left|\boldsymbol{x}\right\rangle \left\langle \boldsymbol{x}\right| \hat{U}_{\boldsymbol{\phi}}^{\dagger}. \tag{2.1}$$

Here, $\boldsymbol{x} \in \{0,1\}^n$ denotes an arbitrary element in the set of all bitstrings of length n corresponding to the computational basis states $\left|\boldsymbol{x}\right\rangle$ of $\mathcal{H}$. The parameters $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\phi}) \in \mathbb{R}^d$ specify the classical probability distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x})$, which captures classical correlations, and the unitary quantum neural network (QNN) [5] $U_{\boldsymbol{\phi}}$, which adds quantum correlations to our representation. We specifically consider $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ to be given by a classical energy-based model (EBM) [12, 7],

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}, \;\; Z_{\boldsymbol{\theta}} = \sum_{\boldsymbol{x}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}, \tag{2.2}$$

where the probability of a given sample $\boldsymbol{x}$ is proportional to the exponential of the energy function $E_{\boldsymbol{\theta}}(\boldsymbol{x})$, with the normalization factor being the partition function $Z_{\boldsymbol{\theta}}$. This definition

gives rise to the equivalent exponential representation,

$$\hat{\rho}_{\boldsymbol{\Omega}} = \frac{1}{Z_{\boldsymbol{\theta}}} e^{-\hat{K}_{\boldsymbol{\Omega}}}, \;\; \hat{K}_{\boldsymbol{\Omega}} = \hat{U}_{\boldsymbol{\phi}} \hat{K}_{\boldsymbol{\theta}} \hat{U}_{\boldsymbol{\phi}}^{\dagger}, \;\; \hat{K}_{\boldsymbol{\theta}} = \sum_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \left|\boldsymbol{x}\right\rangle \left\langle \boldsymbol{x}\right|, \tag{2.3}$$

in which $\hat{\rho}_{\boldsymbol{\Omega}}$ is expressed as the thermal state of a parameterized modular Hamiltonian $\hat{K}_{\boldsymbol{\Omega}}$. Quantum-Hamiltonian-based models (QHBMs) can therefore be seen a a direct quantum generalization of classical energy-based models (EBMs).

The exponential structure of QHBMs further makes these class of models particularly conducive to tasks that involve optimizing the quantum relative entropy [30], which, for a given pair of density operators $\hat{\rho}, \hat{\sigma} \in \mathcal{M}^{(N)}$, is defined as

$$D(\hat{\rho} \| \hat{\sigma}) = \operatorname{tr}[\hat{\rho}(\log \hat{\rho} - \log \hat{\sigma})] = -S(\hat{\rho}) - \operatorname{tr}[\hat{\rho} \log \hat{\sigma}], \tag{2.4}$$

in terms of the von Neumann entropy $S(\hat{\rho}) = -\operatorname{tr}[\hat{\rho} \log \hat{\rho}]$. The quantum relative entropy is a non-commutative generalization [28] of the Kullback-Leibler divergence [15] which is commonly used as a loss function in classical probabilistic machine learning [9]. We highlight two important properties of the quantum relative entropy. First, $D(\hat{\rho} \| \hat{\sigma}) \geq 0$, which is satisfied with equality if and only if $\hat{\rho} = \hat{\sigma}$. Consequently, we can construct a variational principle where we define the quantum relative entropy between our QHBM representation $\hat{\rho}_{\boldsymbol{\Omega}}$ and some desired target state $\hat{\sigma}$ as our loss function, which we then minimize to find optimal parameters $\boldsymbol{\Omega}^*$ such that $\hat{\rho}_{\boldsymbol{\Omega}^*} \approx \hat{\sigma}$. Furthermore, $D(\hat{\rho} \| \hat{\sigma})$ is asymmetric with respect to its arguments $\hat{\rho}$ and $\hat{\sigma}$. By considering both possible orderings of $\hat{\rho}_{\boldsymbol{\Omega}}$ and $\hat{\sigma}$, we obtain two distinct sets of applications for QHBMs, namely variational quantum thermalization (VQT) and quantum modular Hamiltonian learning (QMHL).

2.2 Variational Quantum Thermalization

Suppose we are given a Hamiltonian H and a inverse temperature β and the task is to simulate the associated thermal state

$$\hat{\sigma}_{\beta} = \frac{1}{Z_{\beta}} e^{-\beta \hat{H}}, \;\; Z_{\beta} = \operatorname{tr}\left[e^{-\beta \hat{H}}\right]. \tag{2.5}$$

We may formulate this quantum simulation task as a specific quantum-probabilistic optimization problem, which we term quantum variational thermalization (VQT), by minimizing the forward quantum relative entropy between our QHBM and the target thermal state,

$$\begin{aligned} \min_{\boldsymbol{\Omega}} D(\hat{\rho}_{\boldsymbol{\Omega}} \| \hat{\sigma}_{\beta}) &= \min_{\boldsymbol{\Omega}} \left[\beta \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}} \hat{H}\right] - S(\hat{\rho}_{\boldsymbol{\Omega}}) + \log Z_{\beta}\right] \\ &= \min_{\boldsymbol{\Omega}} \left[\beta \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}} \hat{H}\right] - S(p_{\boldsymbol{\theta}})\right] \\ &= \min_{\boldsymbol{\Omega}} \mathcal{L}_{\mathrm{VQT}}(\boldsymbol{\Omega}). \end{aligned} \tag{2.6}$$

In the first line we have substituted the form of $\hat{\sigma}_\beta$ (2.5) into the definition of the quantum relative entropy (2.4), while in second line we have recognized that the von Neumann entropy of $\hat{\rho}_{\boldsymbol{\Omega}}$ is simply the classical entropy of $p_{\boldsymbol{\theta}}$ and further that $\log Z_\beta$ is independent of $\boldsymbol{\Omega}$. We see that the VQT loss $\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega})$ is directly proportional to the free energy of the state $\hat{\rho}_{\boldsymbol{\Omega}}$ with respect to the system H, $F(\hat{\rho}_{\boldsymbol{\Omega}}) = \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}}\hat{H}\right] - \frac{1}{\beta}S(p_{\boldsymbol{\theta}})$. To optimize the objective (2.6), we may utilize gradient-based methods, such as gradient descent, which are generally preferred over gradient-free optimization. In particular, we find that QHBMs naturally lend themselves to analytical expressions for the gradient of quantum relative entropies that involve both classical and quantum expectations and can therefore be conveniently approximated through sample-based estimates. In the specific case of the VQT loss (2.6), it can be shown that the gradient with respect to the classical probabilistic parameters $\boldsymbol{\theta}$ is [29]

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega}) &= \beta\nabla_{\boldsymbol{\theta}}\operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}}\hat{H}\right] - \nabla_{\boldsymbol{\theta}}S\left(p_{\boldsymbol{\theta}}\right) \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\beta\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle} - E_{\boldsymbol{\theta}}(\boldsymbol{x})\right]\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})\right] \\ &\quad - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\left(\beta\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle} - E_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})\right]\end{aligned} \tag{2.7}$$

where $\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle} = \langle\boldsymbol{x}|\, U_{\boldsymbol{\phi}}^\dagger\hat{H}U_{\boldsymbol{\phi}}\,|\boldsymbol{x}\rangle$ is the expectation of $\hat{H}$ in the state $\hat{U}_{\boldsymbol{\phi}}\,|\boldsymbol{x}\rangle$. Morever, the gradient with respect to the QNN parameters $\boldsymbol{\phi}$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega}) &= \beta\nabla_{\boldsymbol{\phi}}\operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}}\hat{H}\right] - \nabla_{\boldsymbol{\phi}}S\left(p_{\boldsymbol{\theta}}\right) \\ &= \beta\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\nabla_{\boldsymbol{\phi}}\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle}\right].\end{aligned} \tag{2.8}$$

We recognize that $\nabla_{\boldsymbol{\phi}}\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle}$ is the gradient of the expectation of a quantum observable with respect to a parameterized state, a typical scenario encountered in training QNNs through gradient-based optimization. There exist various methods to obtains such gradients, however, the standard approach is to leverage parameter-shift rules [5], which analytically express each partial derivative as a linear combination of parameter-shifted expectation values. For instance, in the case where the QNN component of our QHBM is given by the hardware-efficient ansatz (i.e. a QNN whose parameterized operations are independently parameterized and are of the form of simple exponentials of single Pauli operators, e.g. $U_{\boldsymbol{\phi}} = \prod_j e^{i\phi_j\hat{P}_j}$), we have the following formula for each element of the gradient $\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega})$:

$$\begin{aligned}\partial_{\phi_j}\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega}) =& \beta\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\partial_{\phi_j}\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}}|\boldsymbol{x}\rangle}\right] \\ =& \frac{\beta}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}+\boldsymbol{\Delta}^j}|\boldsymbol{x}\rangle} - \langle\hat{H}\rangle_{\hat{U}_{\boldsymbol{\phi}-\boldsymbol{\Delta}^j}|\boldsymbol{x}\rangle}\right]\end{aligned} \tag{2.9}$$

with $\boldsymbol{\Delta}^j$ being a $\pi/2$ magnitude shift in the j^{th} parameter ϕ_j such that $\boldsymbol{\Delta}^j_k = \frac{\pi}{2}\delta_{jk}$.

2.3 Quantum Modular Hamiltonian Learning

Dual to VQT, suppose we are given query access to an otherwise unknown target state $\hat{\sigma}$ for which seek to learn a representation of in the form of a QHBM. Now taking the reverse direction of the quantum relative entropy as our objective, we obtain quantum modular Hamiltonian learning (QMHL),

$$\begin{aligned}\min_{\boldsymbol{\Omega}} D(\hat{\sigma}\|\hat{\rho}_{\boldsymbol{\Omega}}) &= \min_{\boldsymbol{\Omega}} \left[\operatorname{tr}\left[\hat{\sigma}\hat{K}_{\boldsymbol{\Omega}}\right] - S(\hat{\sigma}) + \log Z_{\boldsymbol{\theta}}\right] \\ &= \min_{\boldsymbol{\Omega}} \left[\operatorname{tr}\left[\hat{\sigma}\hat{K}_{\boldsymbol{\Omega}}\right] + \log Z_{\boldsymbol{\theta}}\right] \\ &= \min_{\boldsymbol{\Omega}} \mathcal{L}_{\text{QMHL}}(\boldsymbol{\Omega}),\end{aligned} \tag{2.10}$$

where we have applied the exponential QHBM representation (2.3) in the first step and identified the independence of $S(\hat{\sigma})$ with respect to $\boldsymbol{\Omega}$ in the second step. The resulting QMHL loss is equivalent to the quantum cross entropy between the target state and the QHBM, $S(\hat{\sigma}, \hat{\rho}_{\boldsymbol{\Omega}}) = -\operatorname{tr}[\hat{\sigma}\log\hat{\rho}_{\boldsymbol{\Omega}}]$. We find that the gradient of the loss with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{QMHL}}(\boldsymbol{\Omega}) &= \nabla_{\boldsymbol{\theta}} \operatorname{tr}\left[\hat{\sigma}\hat{K}_{\boldsymbol{\Omega}}\right] + \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \\ &= \mathbb{E}_{\boldsymbol{x}\sim\sigma_{\boldsymbol{\phi}}(\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p_{\theta}(\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\right]\end{aligned} \tag{2.11}$$

where $\sigma_{\boldsymbol{\phi}}(\boldsymbol{x}) = \langle\boldsymbol{x}|\,\hat{U}_{\boldsymbol{\phi}}^{\dagger}\hat{\sigma}\hat{U}_{\boldsymbol{\phi}}\,|\boldsymbol{x}\rangle$ is the distribution induced by sampling the state $\hat{\sigma}_{\boldsymbol{\phi}} = \hat{U}_{\boldsymbol{\phi}}^{\dagger}\hat{\sigma}\hat{U}_{\boldsymbol{\phi}}$ in the computational basis. The gradient with respect to $\boldsymbol{\phi}$ is shown to be

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\mathcal{L}_{\text{QMHL}}(\boldsymbol{\Omega}) &= \nabla_{\boldsymbol{\phi}} \operatorname{tr}\left[\hat{\sigma}\hat{K}_{\boldsymbol{\Omega}}\right] + \nabla_{\boldsymbol{\phi}} \log Z_{\boldsymbol{\theta}} \\ &= \nabla_{\boldsymbol{\phi}}\langle\hat{K}_{\boldsymbol{\theta}}\rangle_{\hat{\sigma}_{\boldsymbol{\phi}}}\end{aligned} \tag{2.12}$$

where $\langle\hat{K}_{\boldsymbol{\theta}}\rangle_{\hat{\sigma}_{\boldsymbol{\phi}}} = \operatorname{tr}\left[\hat{\sigma}_{\boldsymbol{\phi}}\hat{K}_{\boldsymbol{\theta}}\right]$ is the expectation of $\hat{K}_{\boldsymbol{\theta}}$ in the state $\hat{\sigma}_{\boldsymbol{\phi}}$. Assuming the hardware-efficient ansatz and applying the corresponding parameter-shift rule, we can write this gradient element-wise as

$$\partial_{\phi_j}\mathcal{L}_{\text{QMHL}}(\boldsymbol{\Omega}) = \partial_{\phi_j} \operatorname{tr}\left[\hat{\sigma}_{\boldsymbol{\phi}}\hat{K}_{\boldsymbol{\theta}}\right] \tag{2.13}$$

$$= \frac{1}{2}\left(\langle\hat{K}_{\boldsymbol{\theta}}\rangle_{\hat{\sigma}_{\boldsymbol{\phi}+\boldsymbol{\Delta}^j}} - \langle\hat{K}_{\boldsymbol{\theta}}\rangle_{\hat{\sigma}_{\boldsymbol{\phi}-\boldsymbol{\Delta}^j}}\right). \tag{2.14}$$

Chapter 3

Theory

3.1 Quantum-Probabilistic Natural Gradient Descent

We motivate the utility of leveraging the information geometry of density operators for variational optimization of QHBMs by first recognizing the standard method of vanilla gradient descent, which is the main optimization algorithm used in the prior work [29], as the direction of steepest descent with respect to the Euclidean geometry of the classical parameter space. Formally, we have

$$\boldsymbol{\delta}_{j+1} = \underset{\frac{1}{2}\|\boldsymbol{\delta}\|_2^2 \leq \varepsilon^2}{\arg\min}\ \mathcal{L}(\boldsymbol{\Omega_j} + \boldsymbol{\delta}). \tag{3.1}$$

Here, the loss $\mathcal{L}(\boldsymbol{\Omega}_j)$ is taken to be the quantum relative entropy between the current QHBM representation $\hat{\rho}_{\boldsymbol{\Omega}_j}$ and the target state $\hat{\sigma}$ in either the forward direction $D(\hat{\rho}_{\boldsymbol{\Omega}_j}\|\hat{\sigma})$, which corresponds to the VQT loss $\mathcal{L}_{\text{VQT}}(\boldsymbol{\Omega}_j)$ (2.6), or the reverse direction $D(\hat{\sigma}\|\hat{\rho}_{\boldsymbol{\Omega}_j})$, which corresponds to the QMHL loss $\mathcal{L}_{\text{QMHL}}(\boldsymbol{\Omega}_j)$ (2.10). Furthermore, $\boldsymbol{\delta}_{j+1} = \boldsymbol{\Omega}_{j+1} - \boldsymbol{\Omega}_j$ is the update to the parameters $\boldsymbol{\Omega}$ of the QHBM at iteration $j+1$, and ε is a constant defining the effective step size of our update. We construct the relaxed Lagrangian corresponding to this constrained optimization problem, expand the value of $\mathcal{L}$ to first order in $\boldsymbol{\delta}$, and remove constants that do not depend on $\boldsymbol{\delta}$ to obtain the following unconstrained optimization problem:

$$\boldsymbol{\delta}_{j+1} = \underset{\boldsymbol{\delta}}{\arg\min}\left[\left\langle \nabla_{\boldsymbol{\Omega}_j}\mathcal{L}(\hat{\rho}_{\boldsymbol{\Omega}_j}), \boldsymbol{\delta}\right\rangle + \frac{\lambda}{2}\|\boldsymbol{\delta}\|_2^2\right]. \tag{3.2}$$

Applying the first order optimality condition and rearranging terms produces the familiar update rule of vanilla gradient descent,

$$\boldsymbol{\Omega}_{j+1} = \boldsymbol{\Omega}_j - \frac{1}{\lambda}\nabla_{\boldsymbol{\Omega}_j}\mathcal{L}(\hat{\rho}_{\boldsymbol{\Omega}_j}), \tag{3.3}$$

where we identify $1/\lambda$ as the learning rate. This interpretation highlights the strong dependence of vanilla gradient descent on the Euclidean geometry of parameter space, which is tied to the specific parameterization of our model.

It is instead more natural to perform steepest descent in the distribution space of density operators by using a more appropriate distinguisability measure, which we shall take to be the quantum relative entropy,

$$\boldsymbol{\delta}_{j+1} = \underset{\boldsymbol{\delta}:\ D(\hat{\rho}_{\boldsymbol{\Omega}_j} \| \hat{\rho}_{\boldsymbol{\Omega}_j + \boldsymbol{\delta}}) \leq \varepsilon^2}{\arg\min} \mathcal{L}(\boldsymbol{\Omega_j} + \boldsymbol{\delta}). \tag{3.4}$$

There are two basic conceptual reasons why this formulation of natural gradient or metric-aware descent may be particularly advantageous as compared to vanilla gradient descent:

Idea 1 By taking constant steps over our intended search space of density operators as opposed to the classically parameterizing space, we can diminish [16], and, in some cases, eliminate [1], dependencies on the choice of parameterization.

Idea 2 As the quantum relative entropy is directly used as the variational loss function for QHBMs, these constant steps in quantum relative entropy respect the fundamental distinguishibility of density operators associated with the loss.

To derive the corresponding update rule, we recognize the second-order Taylor series approximation of $D(\hat{\rho}_{\boldsymbol{\Omega}_j} \| \hat{\rho}_{\boldsymbol{\Omega}_j + \boldsymbol{\delta}})$ to be given by

$$D(\hat{\rho}_{\boldsymbol{\Omega}_j} \| \hat{\rho}_{\boldsymbol{\Omega}_j + \boldsymbol{\delta}}) \approx \frac{1}{2} \left\langle \boldsymbol{\delta}, \mathcal{I}(\boldsymbol{\Omega}_j) \boldsymbol{\delta} \right\rangle . \tag{3.5}$$

We note that the first two orders of expansion vanish since $D(\hat{\rho}_{\boldsymbol{\Omega}_j} \| \hat{\rho}_{\boldsymbol{\Omega}_j + \boldsymbol{\delta}})$ is at a minimum of zero when $\boldsymbol{\delta} = 0$. Here, we also leverage the fact that the Hessian of the quantum relative entropy is given by the Bogoliubov-Kubo-Mori (BKM) information matrix, which resolves the BKM metric tensor defined on the manifold of density operators $\mathcal{M}^{(N)}$ to the coordinate basis of the classical parameters $\boldsymbol{\Omega}$. We provide analytical expressions for the elements of the BKM information matrix conducive to sample-based approximation in Appendix A.1. Intuitively, we may interpret $\mathcal{I}$ as providing a notion of how changes in parameter space induce corresponding changes in state space. We further remark that the quantum relative entropy is symmetric up to second-order in $\boldsymbol{\delta}$ such that we could have equivalently employed the opposite direction $D(\hat{\rho}_{\boldsymbol{\Omega}_j + \boldsymbol{\delta}} \| \hat{\rho}_{\boldsymbol{\Omega}_j})$ in our construction and still obtain the same result. Proceeding in an analogous manner as before and expanding all terms to first non-vanishing order in $\boldsymbol{\delta}$, we arrive at

$$\boldsymbol{\delta}_{j+1} = \arg\min_{\boldsymbol{\delta}} \left[\left\langle \nabla_{\boldsymbol{\Omega}_j} \mathcal{L}(\hat{\rho}_{\boldsymbol{\Omega}_j}), \boldsymbol{\delta} \right\rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \mathcal{I}(\boldsymbol{\Omega}_j) \boldsymbol{\delta} \right\rangle \right] . \tag{3.6}$$

The solution to the above optimization problem yields the quantum-probabilistic natural gradient descent (QPNGD) update rule,

$$\boldsymbol{\Omega}_{j+1} = \boldsymbol{\Omega}_j - \frac{1}{\lambda} \mathcal{I}^{-1}(\boldsymbol{\Omega}_j) \nabla_{\boldsymbol{\Omega}_j} \mathcal{L}(\boldsymbol{\Omega}_j). \tag{3.7}$$

Algorithm 1 Quantum-Probabilistic Natural Gradient Descent (QPNGD)

for $j = 1, 2, \ldots$ **do**
  select λ_j
  evaluate $\nabla_{\mathbf{\Omega}_j} \mathcal{L}(\mathbf{\Omega}_j)$
  construct $\mathcal{I}(\mathbf{\Omega}_j)$ via (15), (18), and (19)
  compute $\mathcal{I}^{-1}(\mathbf{\Omega}_j)$
  update $\mathbf{\Omega}_{j+1} \leftarrow \mathbf{\Omega}_j - \frac{1}{\lambda_j} \mathcal{I}^{-1}(\mathbf{\Omega}_j) \nabla \mathcal{L}(\mathbf{\Omega}_j)$

We present the full procedure in Algorithm 1. Indeed, this aesthetically matches the classical natural gradient update rule [2] and existing quantum generalizations [25, 14, 26]. In the classical case, the distribution space is that of categorical probability distributions and the distinguishability measure is the classical relative entropy, which gives rise to the Fisher-Rao metric and the classical Fisher information matrix. In the quantum case, however, we note that there exists a degeneracy in the choice of metric over the space of density operators, and existing works have thus far considered the Bures-Helstrom metric and the corresponding quantum Fisher information matrix. We alternatively choose to utilize the Bogoliubov-Kubo-Mori (BKM) information geometry for the following reasons that are both first-known for quantum metric-aware descent rules:

Contribution 1 The QPNGD update rule (3.44) admits a provable optimality guarantee in terms of the variance of the estimated parameters.

Contribution 2 The QPNGD update rule (3.44) admits a tractable dual form conducive to sample-efficiency improvements.

3.2 Quantum Fisher Efficiency

We describe a notion of optimality which we may seek to satisfy for a general optimization algorithm. In particular, we consider the idea of Fisher efficiency, and describe its quantum analogue, quantum Fisher efficiency, both of which take the perspective of viewing iterative update rules as parameter estimation strategies. In the classical case, attaining Fisher efficiency means that the asymptotic accuracy of an unbiased estimator, as measured by its error covariance matrix, attains the well-known classical Cramér-Rao bound to first-order in the number of data samples used. We may think of an online optimization rule as a statistical estimator by saying that the latest parameters at step j is the estimator given $O(j)$ data samples. Fisher efficiency is achieved for classical online natural gradient descent with a particular choice of learning rate, assuming that the optimal parameters are eventually

reached [1]. This therefore implies that, to first-order, such an update rule can achieve the best-case asymptotic scaling that is usually associated with maximum likelihood estimation.

Analogously, an unbiased estimator is said to achieve quantum Fisher efficiency if it saturates the generalized quantum Cramér-Rao bound, which, in specific case of the BKM geometry, takes the form

$$\mathrm{Cov}(\hat{\boldsymbol{A}}_j; \boldsymbol{\Omega}^*) \succeq \frac{1}{j}\mathcal{I}^{-1}(\boldsymbol{\Omega}_j). \tag{3.8}$$

Here, $\hat{\boldsymbol{A}}_j = \{\hat{A}_0^k\}_{k=1}^d$ is a collection of quantum observables that satisfy $\mathrm{tr}\left[\hat{\rho}_{\boldsymbol{\Omega}^*}\hat{A}_j^k\right] = \Omega_j^k$, where Ω_j^k denotes the k^{th} element of $\boldsymbol{\Omega}_j$, so that $\hat{\boldsymbol{A}}_j$ is an unbiased estimator of $\boldsymbol{\Omega}_j$, and $\mathrm{Cov}(\hat{\boldsymbol{A}}_j; \boldsymbol{\Omega}^*)$ is the error covariance matrix of $\hat{\boldsymbol{A}}_j$ relative to the optimal parameters $\boldsymbol{\Omega}^*$. No known such result has been shown for prior constructions of quantum metric-aware descent rules [25, 14, 26]. The intuitive reason is due the fundamental discrepancy between the chosen metric, which captures the curvature of the state space, and the curvature of the objective function. We demonstrate that our particular quantum metric-aware descent rule is the first known to provide such a guarantee. This is because we are able to take advantage of the fact that the variational loss for the QHBM class of models is precisely given by our selected distinguishability measure of the quantum relative entropy so that the BKM information matrix is the Hessian of the loss.

We specifically consider an online version of the QPNGD update rule with a particular choice of learning rate,

$$\boldsymbol{\Omega}_{j+1} = \boldsymbol{\Omega}_j - \frac{1}{j}\mathcal{I}^{-1}(\boldsymbol{\Omega}_j)\tilde{\nabla}_{\boldsymbol{\Omega}_j}\mathcal{L}(\boldsymbol{\Omega}_j). \tag{3.9}$$

Here, our target state is taken to be the QHBM representation generated by the optimal parameters $\hat{\rho}_{\boldsymbol{\Omega}^*}$, which implies that the loss is accordingly $\mathcal{L}(\boldsymbol{\Omega}_j) = D(\hat{\rho}_{\boldsymbol{\Omega}_j}\|\hat{\rho}_{\boldsymbol{\Omega}^*})$ or $\mathcal{L}(\boldsymbol{\Omega}_j) = D(\hat{\rho}_{\boldsymbol{\Omega}^*}\|\hat{\rho}_{\boldsymbol{\Omega}_j})$. Morever, $\tilde{\nabla}_{\boldsymbol{\Omega}_j}\mathcal{L}(\boldsymbol{\Omega}_j)$ is an online unbiased estimator of the gradient of the loss obtained by drawing a single pure state $\hat{U}_{\boldsymbol{\phi}^*}|\boldsymbol{x}\rangle$ from the eigenstates of $\hat{\rho}_{\boldsymbol{\Omega}^*}$ (with probability of the corresponding eigenvalue $p_{\boldsymbol{\theta}^*}(\boldsymbol{x})$) at each optimization step such that

$$\mathbb{E}_{\boldsymbol{x}}[\tilde{\nabla}_{\boldsymbol{\Omega}_j}\mathcal{L}(\boldsymbol{\Omega}_j)] = \nabla_{\boldsymbol{\Omega}_j}\mathcal{L}(\boldsymbol{\Omega}_j). \tag{3.10}$$

We claim this learning rule is optimal in the sense of quantum Fisher efficiency, as stated in the following theorem:

Theorem 3.2.1. *Suppose that $\mathcal{I}(\boldsymbol{\Omega})$ is non-singular for all $\boldsymbol{\Omega}$. Suppose further that $\boldsymbol{\Omega}_j$ converges to the optimal parameters $\boldsymbol{\Omega}^*$ in expectation, i.e., $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\Omega}_j] \to \boldsymbol{\Omega}^*$ as $j \to \infty$. In such a case, the online quantum-probabilistic natural gradient descent (QPNGD) update rule (3.9) is quantum Fisher efficient, attaining the quantum Cramér-Rao bound (3.8) asymptotically.*

Proof. We prove the result of Theorem 3.2.1 by mapping the online QPNGD update rule (3.9) in parameter space to a latent dynamical equation in quantum observables so that

we may be able to utilize the parameter estimatation language of the generalized quantum Cramér-Rao bound (3.8). To be explicit, we take our particular loss function to be the reverse quantum relative entropy $\mathcal{L}(\mathbf{\Omega}_j) = D(\hat{\rho}_{\mathbf{\Omega}^*} \| \hat{\rho}_{\mathbf{\Omega}_j})$, however, we note that the same result will still hold if instead we have the forward direction, as both directions are symmetric up to second-order. Using the collection of quantum observables $\hat{\boldsymbol{A}}_j$ previously defined in the context of (3.8), we rewrite (3.9) element-wise in expectation as

$$\mathrm{tr}\left[\hat{\rho}_{\mathbf{\Omega}^*} \hat{A}^k_{j+1}\right] = \mathrm{tr}\left[\hat{\rho}_{\mathbf{\Omega}^*} \hat{A}^k_j\right] - \frac{1}{j} \sum_l [\mathcal{I}^{-1}(\mathbf{\Omega}_j)]_{k,l} \partial_{\mathbf{\Omega}^l_j} D(\hat{\rho}_{\mathbf{\Omega}^*} \| \hat{\rho}_{\mathbf{\Omega}_j}). \tag{3.11}$$

We note that for the BKM geometry, we can relate the Hessian of the quantum relative entropy to the definition of the information matrix,

$$-\partial^2_{\mathbf{\Omega}'_l, \mathbf{\Omega}_m} D(\hat{\rho}_{\mathbf{\Omega}'} \| \hat{\rho}_{\mathbf{\Omega}}) \mid_{\mathbf{\Omega}'=\mathbf{\Omega}} = [\mathcal{I}(\mathbf{\Omega})]_{l,m} = \mathrm{tr}[(\partial_{\mathbf{\Omega}_l} \hat{\rho}_{\mathbf{\Omega}}) \mathscr{L}_{\mathbf{\Omega}} (\partial_{\mathbf{\Omega}_m} \hat{\rho}_{\mathbf{\Omega}})]. \tag{3.12}$$

Here, we define the raising operator $\mathscr{R}_{\mathbf{\Omega}}$ as playing the role of the metric with raised indices, which in the case of the BKM metric is given by

$$\mathscr{R}_{\mathbf{\Omega}}(\hat{A}) = \int_0^1 \hat{\rho}^s_{\mathbf{\Omega}} \hat{A} \hat{\rho}^{1-s}_{\mathbf{\Omega}} ds, \tag{3.13}$$

whereas the corresponding lowering operator $\mathscr{L}_{\mathbf{\Omega}} = \mathscr{R}^{-1}_{\mathbf{\Omega}}$ acts as the metric with lowered indices and satisfies

$$\mathscr{L}_{\mathbf{\Omega}}(\hat{A}) = \int_0^\infty (\hat{\rho}_{\mathbf{\Omega}} + s\mathbb{1})^{-1} \hat{A} (\hat{\rho}_{\mathbf{\Omega}} + s\mathbb{1})^{-1} ds. \tag{3.14}$$

It follows from (3.12) that the first derivative of the quantum relative entropy can be expressed as

$$-\partial_{\mathbf{\Omega}_l} D(\hat{\rho}_{\mathbf{\Omega}'} \| \hat{\rho}_{\mathbf{\Omega}}) = \mathrm{tr}[\hat{\rho}_{\mathbf{\Omega}'} \mathscr{L}_{\mathbf{\Omega}} (\partial_{\mathbf{\Omega}_l} \hat{\rho}_{\mathbf{\Omega}})]. \tag{3.15}$$

Given (3.15), we obtain

$$\mathrm{tr}\left[\hat{\rho}_{\mathbf{\Omega}^*} \hat{A}^k_{j+1}\right] = \mathrm{tr}\left[\hat{\rho}_{\mathbf{\Omega}^*} \hat{A}^k_j\right] + \frac{1}{j} \sum_l [\mathcal{I}^{-1}(\mathbf{\Omega}_j)]_{k,l} \,\mathrm{tr}\left[\hat{\rho}_{\mathbf{\Omega}^*} \mathscr{L}_{\mathbf{\Omega}_j} (\partial_{\mathbf{\Omega}^l_j} \hat{\rho}_{\mathbf{\Omega}_j})\right]. \tag{3.16}$$

We see that this is simply the quantum expectation of the latent dynamical equation

$$\hat{A}^k_{j+1} = \hat{A}^k_j + \frac{1}{j} \sum_l [\mathcal{I}^{-1}(\mathbf{\Omega}_j)]_{k,l} \mathscr{L}_{\mathbf{\Omega}_j} (\partial_{\mathbf{\Omega}^l_j} \hat{\rho}_{\mathbf{\Omega}_j}), \tag{3.17}$$

which describes how our collection of quantum observables, acting as estimators of the classical parameters, updates at each descent step. Without loss of generality, we assume that

$\mathbf{\Omega}^* = 0$ and define the associated error covariance matrix at iteration j as given element-wise by

$$[V_j]_{k,l} = [\mathrm{Cov}(\hat{\boldsymbol{A}}_j; \mathbf{\Omega}^*)]_{k,l} = \mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}^*}\left(\hat{A}_j^k\right)\hat{A}_j^l\right]. \tag{3.18}$$

We recognize that V_j is indeed symmetric by the property $\mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}}(\hat{A})\hat{B}\right] = \mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}}(\hat{B}^\dagger)\hat{A}^\dagger\right]$ [22]. Direction substitution of (3.17) induces a corresponding dynamical equation on the elements of the error covariance matrices,

$$\begin{aligned}[V_{j+1}]_{k,l} = [V_j]_{k,l} + \frac{2}{j}\sum_{l'}[\mathcal{I}^{-1}(\mathbf{\Omega}_j)]_{l,l'}\,\mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}^*}(\hat{A}_j^k)\mathscr{L}_{\mathbf{\Omega}_j}\left(\partial_{\mathbf{\Omega}_j^{l'}}\hat{\rho}_{\mathbf{\Omega}_j}\right)\right] \\ + \frac{1}{j^2}[\mathcal{I}^{-1}(\mathbf{\Omega}_j)]_{k,l} + O\left(\frac{1}{j^3}\right),\end{aligned} \tag{3.19}$$

where we have used

$$\mathscr{R}_{\mathbf{\Omega}^*}(\hat{A}) = \mathscr{R}_{\mathbf{\Omega}_j}(\hat{A}) + O\left(\frac{1}{j}\right), \tag{3.20}$$

since $\mathbf{\Omega}_j$ converges to $\mathbf{\Omega}^*$, along with (3.12) and the fact that $\mathscr{L}_{\mathbf{\Omega}} = \mathscr{R}_{\mathbf{\Omega}}^{-1}$. Taking the gradient of the second-order Taylor series approximation of $D(\hat{\rho}_{\mathbf{\Omega}^*}\|\hat{\rho}_{\mathbf{\Omega}_j})$,

$$D(\hat{\rho}_{\mathbf{\Omega}^*}\|\hat{\rho}_{\mathbf{\Omega}_j}) \approx \frac{1}{2}\left\langle \mathbf{\Omega}_j - \mathbf{\Omega}^*, \mathcal{I}(\mathbf{\Omega}^*)(\mathbf{\Omega}_j - \mathbf{\Omega}^*)\right\rangle, \tag{3.21}$$

we obtain

$$\nabla_{\mathbf{\Omega}_j} D(\hat{\rho}_{\mathbf{\Omega}^*}\|\hat{\rho}_{\mathbf{\Omega}_j}) \approx \left\langle \mathcal{I}(\mathbf{\Omega}^*), \mathbf{\Omega}_j - \mathbf{\Omega}^*\right\rangle. \tag{3.22}$$

We then apply this result in conjunction with and (3.15) to yield

$$\mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}^*}(\hat{A}_j^k)\mathscr{L}_{\mathbf{\Omega}_j}\left(\partial_{\mathbf{\Omega}_j^{l'}}\hat{\rho}_{\mathbf{\Omega}_j}\right)\right] = -\sum_m [\mathcal{I}(\mathbf{\Omega}^*)]_{l',m}\,\mathrm{tr}\left[\mathscr{R}_{\mathbf{\Omega}^*}(\hat{A}_j^k)\hat{A}_j^m\right] + O\left(\frac{1}{j^2}\right) \tag{3.23}$$

$$= -[\mathcal{I}(\mathbf{\Omega}^*)V_j]_{l',k} + O\left(\frac{1}{j^2}\right) \tag{3.24}$$

Substituting (3.23) into (3.19) and expressing the result in matrix form gives

$$V_{j+1} = V_j - \frac{2}{j}\mathcal{I}^{-1}(\mathbf{\Omega}_j)\mathcal{I}(\mathbf{\Omega}^*)V_j + \frac{1}{j^2}\mathcal{I}^{-1}(\mathbf{\Omega}_j) + O\left(\frac{1}{j^3}\right). \tag{3.25}$$

Noting that

$$\mathcal{I}^{-1}(\mathbf{\Omega}_j) = \mathcal{I}^{-1}(\mathbf{\Omega}^*) + O\left(\frac{1}{j}\right), \tag{3.26}$$

we finally arrive at

$$V_{j+1} = V_j - \frac{2}{j}V_j + \frac{1}{j^2}(\mathcal{I}_{\boldsymbol{\Omega}^*})^{-1} + O\left(\frac{1}{j^3}\right). \tag{3.27}$$

The solution to (3.27) is asymptotically

$$V_j = \frac{1}{j}\mathcal{I}^{-1}(\boldsymbol{\Omega}_j), \tag{3.28}$$

which satisfies (3.8) with equality. □

Since the online update rule achieves quantum Fisher efficiency, it is straightforward to see that using more data samples at each optimization step, as would be typical for a standard batch learning rule, can only improve convergence and therefore also achieves quantum Fisher efficiency.

3.3 Quantum-Probabilistic Mirror Descent

Estimating the BKM information matrix, as per (15), (18), and (19), for each application of the QPNGD update rule is clearly a challenging task in terms of both memory and algorithmic complexity given that the number of unique elements, $d(d+1)/2$, is quadratic in the dimensionality of the parameter space $d = |\boldsymbol{\Omega}|$. In particular, if the QNN component of the QHBM is amenable to parameter shift rules and $q = |\boldsymbol{\phi}|$ is specifically the number of parameters of the QNN, then $2q(q+2)$ quantum expectation evaluations are required, which constitutes the dominant factor of computation. Finding tractable approximations for the various metric-aware descent constructions has been recognized as critical for practical application in both quantum [25] and classical [18, 24] works. In the quantum case, block approximations [25] have been considered as mechanisms to limit the computation of cross terms to pairs of parameters which are expected to be significantly correlated. Related types of inductive biases have been successful classically, for example assuming that the information matrix has a Kronecker product factorization [18]. In the classical literature, a separate result is known which translates the second-order method of natural gradient descent to the first-order method of mirror descent [24]. The result follows from the fundamental concept of duality by which two coordinate systems can be considered dual to each other in the sense of being related by the Legendre transform. We now seek to apply this notion to our current construction of QPNGD in order to achieve improvements in quantum sample-efficiency.

There exist two natural choices of coordinate systems on the space of density operators $\mathcal{M}^{(N)}$ [11]. Specifically, mixture or Bloch coordinates $\{\eta_j\}_{j=1}^{N^2-1}$ refer to the decomposition of density matrices in a basis identifiable with $SU(N)$ as

$$\hat{\rho}_{\boldsymbol{\eta}} = \frac{1}{N}\mathbb{1} + \sum_{j=1}^{N^2-1} \eta_j \hat{\sigma}_j, \tag{3.29}$$

where $\hat{\sigma}_j$ are traceless, Hermitian matrices and $\eta_j = \frac{1}{2}\operatorname{tr}[\hat{\sigma}_j \hat{\rho}_{\boldsymbol{\Omega}}]$. The positivity of density matrices implies that the mixture coordinates must be constrained such that $\sqrt{\sum_{j=1}^{N^2-1} \eta_j^2} \leq 1$. We recognize that the mixture representation of QHBMs $\{p_{\boldsymbol{\theta}}, \boldsymbol{\phi}\}$ (2.1) are directly related to the mixture coordinates $\{\eta_j\}_{j=1}^{N^2-1}$, with $p_{\boldsymbol{\theta}}$ likewise being constrained over the reals. Alternatively, exponential coordinates $\{\varphi_j\}_{j=1}^{N^2-1}$ refer to the $SU(N)$-identifiable decomposition given by

$$\hat{\rho}_{\boldsymbol{\varphi}} = \frac{e^{-\hat{K}_{\boldsymbol{\varphi}}}}{Z_{\boldsymbol{\varphi}}}, \;\; \hat{K}_{\boldsymbol{\varphi}} = \sum_{j=1}^{N^2-1} \varphi_j \hat{\sigma}_j, \;\; Z_{\boldsymbol{\varphi}} = \operatorname{tr}\left[e^{-\hat{K}_{\boldsymbol{\varphi}}}\right], \tag{3.30}$$

where we have σ_j as before and $\varphi_j \in \mathbb{R}$ now unconstrained. There analogously exists a relation between the exponential representation $\{E_{\boldsymbol{\theta}}, \boldsymbol{\phi}\}$ (2.3) and the exponential coordinates $\{\varphi_j\}_{j=1}^{N^2-1}$, with $E_{\boldsymbol{\theta}}$ similarly unconstrained over the reals.

It is known that the BKM metric is the unique monotone metric for which the mixture and exponential coordinate systems $(\boldsymbol{\eta}, \boldsymbol{\varphi})$ are mutually dual [11, 10]. Leveraging this property, we now present the following theorem, which posits the equivalence of our quantum-probabilistic formulations of mirror descent and natural gradient descent:

Theorem 3.3.1. *The quantum-probabilistic mirror descent (QPMD) rule in mixture coordinates $\boldsymbol{\eta}$, with the Bregman divergence given by the quantum relative entropy,*

$$\boldsymbol{\eta}_{j+1} = \arg\min_{\boldsymbol{\eta}} \left[\langle \nabla_{\boldsymbol{\eta}_j} \mathcal{L}(\boldsymbol{\eta}_j), \boldsymbol{\eta} \rangle + \lambda D(\hat{\rho}_{\boldsymbol{\eta}} \| \hat{\rho}_{\boldsymbol{\eta}_j})\right], \tag{3.31}$$

is equivalent to the quantum-probabilistic natural gradient descent (QPNGD) rule in exponential coordinates $\boldsymbol{\varphi}$,

$$\boldsymbol{\varphi}_{j+1} = \boldsymbol{\varphi}_j - \frac{1}{\lambda} \mathcal{I}^{-1}(\boldsymbol{\varphi}_j) \nabla_{\boldsymbol{\varphi}_j} \mathcal{L}(\boldsymbol{\varphi}_j). \tag{3.32}$$

A parallel argument holds if the QPMD update rule is expressed in terms of exponential coordinates and the QPNGD update rule in terms of mixture coordinates.

Proof. The duality of the mixture and exponential coordinate systems $(\boldsymbol{\eta}, \boldsymbol{\varphi})$ under the BKM metric implies the existence of a corresponding pair of potential functions $\Phi(\boldsymbol{\eta}), \Psi(\boldsymbol{\varphi})$ that are dual to each other via the Legendre transform,

$$\boldsymbol{\varphi} = \nabla_{\boldsymbol{\eta}} \Phi(\boldsymbol{\eta}), \;\; \boldsymbol{\eta} = \nabla_{\boldsymbol{\varphi}} \Psi(\boldsymbol{\varphi}), \tag{3.33}$$

and further whose Hessians are related to the BKM information matrix [10],

$$\nabla^2_{\boldsymbol{\eta}} \Phi(\boldsymbol{\eta}) = \mathcal{I}(\boldsymbol{\eta}), \;\; \nabla^2_{\boldsymbol{\varphi}} \Psi(\boldsymbol{\varphi}) = \mathcal{I}(\boldsymbol{\varphi}), \tag{3.34}$$

parameterized with respect to the appropriate coordinate system. It can be shown that these functions are the negative von Neumann entropy,

$$\Phi(\boldsymbol{\eta}) = -S(\hat{\rho}_{\boldsymbol{\eta}}) = \mathrm{tr}[\hat{\rho}_{\boldsymbol{\eta}} \log \hat{\rho}_{\boldsymbol{\eta}}], \tag{3.35}$$

and the log partition function [11],

$$\Psi(\boldsymbol{\varphi}) = \log Z_{\boldsymbol{\varphi}} = \log \mathrm{tr}\left[e^{-\hat{K}_{\boldsymbol{\varphi}}}\right], \tag{3.36}$$

where $\rho_{\boldsymbol{\eta}}$ and $\hat{K}_{\boldsymbol{\varphi}}$ are as given in (3.29) and (3.30), respectively. A unique property of the negative von Neumann entropy in particular is that the induced Bregman divergence is precisely the quantum relative entropy [21],

$$D_\Phi(\boldsymbol{\eta}, \boldsymbol{\eta}') = \Phi(\boldsymbol{\eta}) - \Phi(\boldsymbol{\eta}') - \langle \nabla_{\boldsymbol{\eta}'} \Phi(\boldsymbol{\eta}'), \boldsymbol{\eta} - \boldsymbol{\eta}' \rangle = D(\hat{\rho}_{\boldsymbol{\eta}} \| \hat{\rho}_{\boldsymbol{\eta}'}). \tag{3.37}$$

Therefore, applying the first-order optimalilty condition to the minimization problem in (3.31) and using the above fact (3.37), we have

$$\nabla_{\boldsymbol{\eta}_{j+1}} \Phi(\boldsymbol{\eta}_{j+1}) = \nabla_{\boldsymbol{\eta}_j} \Phi(\boldsymbol{\eta}_j) - \frac{1}{\lambda} \nabla_{\boldsymbol{\eta}_j} \mathcal{L}(\boldsymbol{\eta}_j). \tag{3.38}$$

We can rewrite (3.38) in terms of exponential coordinates via the Legendre transform (3.33) to yield

$$\boldsymbol{\varphi}_{j+1} = \boldsymbol{\varphi}_j - \frac{1}{\lambda} \nabla_{\boldsymbol{\eta}_j} \mathcal{L}(\nabla_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j)). \tag{3.39}$$

Recognizing that

$$\nabla_{\boldsymbol{\varphi}_j} \mathcal{L}(\nabla_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j)) = \nabla^2_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j) \nabla_{\boldsymbol{\eta}_j} \mathcal{L}(\nabla_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j)), \tag{3.40}$$

we obtain

$$\nabla_{\boldsymbol{\eta}_j} \mathcal{L}(\nabla_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j)) = \mathcal{I}^{-1}(\boldsymbol{\varphi}_j) \nabla_{\boldsymbol{\varphi}_j} \mathcal{L}(\nabla_{\boldsymbol{\varphi}_j} \Psi(\boldsymbol{\varphi}_j)), \tag{3.41}$$

where we have used the relation $\nabla^2_{\boldsymbol{\varphi}} \Psi(\boldsymbol{\varphi}) = \mathcal{I}(\boldsymbol{\varphi})$ from (3.34). Direction substitution of (3.41) into (3.39) gives the desired result (3.32). □

One consequence of this equivalence is that the optimality guarantee of QPNGD also extends to QPMD such that with the particular choice of regularization $\lambda = j$ and under the same convergence assumptions, the update rule (3.31) is quantum Fisher efficient. In terms of the QHBM parameterization, we may express the QPMD update rule (3.31) as

$$\boldsymbol{\Omega}_{j+1} = \arg\min_{\boldsymbol{\Omega}} \left[\langle \nabla_{\boldsymbol{\Omega}_j} \mathcal{L}(\boldsymbol{\Omega}_j), \boldsymbol{\Omega} \rangle + \lambda D(\hat{\rho}_{\boldsymbol{\Omega}} \| \hat{\rho}_{\boldsymbol{\Omega}_j}) \right]. \tag{3.42}$$

Algorithm 2 Quantum-Probabilistic Mirror Descent (QPMD)

for $j = 1, 2, \ldots$ **do**
  select λ_j
  evaluate $\nabla_{\mathbf{\Omega}_j} \mathcal{L}(\mathbf{\Omega}_j)$
  for $k = 1, 2, \ldots, K$ **do**
    select η_k
    evaluate $\nabla_{\mathbf{\Omega}_j^k} D(\hat{\rho}_{\mathbf{\Omega}_j^k} \| \hat{\rho}_{\mathbf{\Omega}_j})$
    update $\mathbf{\Omega}_j^{k+1} \leftarrow \mathbf{\Omega}_j^k - \eta_k \big(\nabla_{\mathbf{\Omega}_j} \mathcal{L}(\mathbf{\Omega}_j) + \lambda_j \nabla_{\mathbf{\Omega}_j^k} D(\hat{\rho}_{\mathbf{\Omega}_j^k} \| \hat{\rho}_{\mathbf{\Omega}_j}) \big)$
  update $\mathbf{\Omega}_{j+1} \leftarrow \mathbf{\Omega}_j^{K+1}$

By further treating the minimization in (3.42) as a sub-problem that we choose solve with an inner-loop of gradient descent, we obtain Algorithm 2. Though quantum Fisher efficiency is no longer guaranteed to hold under such reparameterization and approximation, in comparison with Algorithm 1, no inversion is required and we have transformed a second-order method to be entirely first-order, implying a potential reduction in quantum sample complexity. Assuming that we perform k steps of gradient descent in the inner-loop to sufficiently converge to the minimum, the number of quantum evaluations at each step is $k(2q+1)$. If k is such that $k(2q+1) < 2q(q+2)$, then QPMD can be utilized as a sample-efficient alternative to QPNGD.

3.4 Learning Sequences of Quantum States

Suppose that instead of a single target state, we now have a sequence of such states $\{\hat{\sigma}(\mathbf{\Lambda}(\tau_k))\}_{k=1}^M$ where $\Lambda : \mathbb{R} \to \mathbb{R}^d$ defines a path in the parameterizing space of target density operators, which for instance may include parameters of time, temperature, or couplings of a given Hamiltonian. Our objective is then to learn a sequence of optimal parameters of a QHBM $\{\mathbf{\Omega}^*(\tau_k)\}_{k=1}^M$ such that the corresponding sequence of generated quantum state representations well approximates each target state, $\hat{\rho}_{\mathbf{\Omega}^*(\tau_k)} \approx \hat{\sigma}(\mathbf{\Lambda}(\tau_k))$. To do so, we can define a loss function for each step in the sequence as the quantum relative entropy, in either direction, between our QHBM and the corresponding target state, and collectively minimize the sum of all losses along the sequence,

$$\{\mathbf{\Omega}^*(\tau_k)\}_{k=1}^M = \underset{\{\mathbf{\Omega}(\tau_k)\}_{k=1}^M}{\arg\min} \sum_{k=1}^M \mathcal{L}(\mathbf{\Omega}(\tau_k)) \tag{3.43}$$

where $\mathcal{L}(\mathbf{\Omega}(\tau_k)) = D(\hat{\rho}_{\mathbf{\Omega}(\tau_k)} \| \hat{\sigma}(\mathbf{\Lambda}(\tau_k)))$ or $\mathcal{L}(\mathbf{\Omega}(\tau_k)) = D(\hat{\sigma}(\mathbf{\Lambda}(\tau_k)) \| \hat{\rho}_{\mathbf{\Omega}(\tau_k)})$. We may naively optimize the objective (3.43) by independently minimizing each loss term in the summation. However, we are then not exploiting our prior knowledge of the information geometry that

adjacent states in the sequence should be close to one another in quantum relative entropy. In particular, if $\hat{\sigma}(\mathbf{\Lambda}(\tau))$ is continuous in trace distance with respect to the parameter τ, then the quantum relative entropy between will likewise be continuous with respect to τ since we are considering a finite-dimensional space of density operators $\mathcal{M}^{(N)}$ [4]. Therefore, if our discretization $\{\tau_k\}_{k=1}^M$ is sufficiently fine, we should expect some notion of locality with respect to the quantum relative entropy to hold. We accordingly describe a natural extension of metric-aware optimization from single states to sequences of states that effectively encodes this prior.

Without loss of generality, we consider the first two states in a given sequence $\{\hat{\sigma}(\mathbf{\Lambda}(\tau_1)), \hat{\sigma}(\mathbf{\Lambda}(\tau_2))\}$ for which we seek to learn a corresponding pair of optimal QHBM parameters $\{\mathbf{\Omega}^*(\tau_1), \mathbf{\Omega}^*(\tau_2)\}$. We may obtain the first set of optimal parameters $\mathbf{\Omega}^*(\tau_1)$ by starting from some random initialization $\mathbf{\Omega}_0(\tau_1)$ and iteratively applying either the QPNGD update rule (3.44) or the QPMD update rule (3.42) on the loss $\mathcal{L}(\mathbf{\Omega}(\tau_1))$ until convergence. Now, if $\hat{\sigma}(\mathbf{\Lambda}(\tau_1))$ is geometrically local to $\hat{\sigma}(\mathbf{\Lambda}(\tau_2))$ in the sense that $D(\hat{\sigma}(\mathbf{\Lambda}(\tau_1))\|\hat{\sigma}(\mathbf{\Lambda}(\tau_2))) \leq \varepsilon^2$ for some small ε, then we reason that the previous optimal parameters may serve as a good initialization for optimizing the next set of parameters with respect to its corresponding loss $\mathcal{L}(\mathbf{\Omega}(\tau_2))$ such that $\mathbf{\Omega}_0(\tau_2) = \mathbf{\Omega}^*(\tau_1)$. Generalizing to any arbitrary step in the sequence, we may link our methods of metric-aware descent as

$$\mathbf{\Omega}_{j+1}(\tau_k) = \mathbf{\Omega}_j(\tau_k) - \frac{1}{\lambda}\mathcal{I}^{-1}(\mathbf{\Omega}_j(\tau_k))\nabla_{\mathbf{\Omega}_j(\tau_k)}\mathcal{L}(\mathbf{\Omega}_j(\tau_k)) \tag{3.44}$$

for the QPNGD update rule (3.44), and

$$\mathbf{\Omega}_{j+1}(\tau_k) = \arg\min_{\mathbf{\Omega}} \left[\langle \mathbf{\Omega}, \nabla\mathcal{L}\left(\mathbf{\Omega}_j(\tau_k)\right)\rangle + \lambda D\left(\hat{\rho}_{\mathbf{\Omega}} \| \hat{\rho}_{\mathbf{\Omega}_j(\tau_k)}\right) \right] \tag{3.45}$$

for the QPMD update rule (3.42), with the initialization $\mathbf{\Omega}_0(\tau_k) = \mathbf{\Omega}^*(\tau_{k-1})$.

Meta-Variational Quantum Thermalization

We can extend VQT to a sequence learning scenario by considering that we are now given a sequence of parameterized Hamiltonians and inverse temperatures $\{\hat{H}_{\mathbf{\Lambda}(\tau_k)}, \beta_k\}_{k=1}^M$ so that our sequence of target states are then the corresponding thermal states $\hat{\sigma}(\mathbf{\Lambda}(\tau_k)) = e^{-\beta_k \hat{H}_{\mathbf{\Lambda}(\tau_k)}}/Z_{\mathbf{\Lambda}(\tau_k)}$. Using the forward direction of the quantum relative entropy as the loss, we arrive at the following sequence learning problem, which we term meta-variational quantum thermalization (meta-VQT):

$$\{\mathbf{\Omega}^*(\tau_k)\}_{k=1}^M = \arg\min_{\{\mathbf{\Omega}(\tau_k)\}} \sum_{k=1}^M D(\hat{\rho}_{\mathbf{\Omega}(\tau_k)} \| \hat{\sigma}(\mathbf{\Lambda}(\tau_k))) \tag{3.46}$$

Meta-Quantum Modular Hamiltonian Learning

Proceeding in an analogous manner for the case of QMHL, assume that we have direct query access to a given sequence of arbitrary target states $\{\hat{\sigma}(\mathbf{\Lambda}(\tau_k))\}_{k=1}^M$. If we accordingly take

the loss to be the reverse direction of the quantum relative entropy, we obtain meta-quantum modular Hamiltonian learning (meta-QMHL),

$$\{\boldsymbol{\Omega}^*(\tau_k)\}_{k=1}^M = \arg\min_{\{\boldsymbol{\Omega}(\tau_k)\}} \sum_{k=1}^M D(\hat{\sigma}(\boldsymbol{\Lambda}(\tau_k))\|\hat{\rho}_{\boldsymbol{\Omega}(\tau_k)}). \tag{3.47}$$

Quantum Variational Recursive Time Evolution Ansatz

Suppose we are alternatively given query access to an initial quantum state $\hat{\sigma}_0$ and the ability to apply a completely-positive trace preserving (CPTP) dynamical map Φ to an arbitrary density operator. Such a map may encode unitary (Schrodinger), Markovian (Lindbladian), or non-Markovian (Nakajima-Zwanzig) dynamics. Our goal is then to simulate the evolution of the initial quantum state under the action of the dynamical map over some time interval $[0, T]$. In particular, we assume that we can discretize the dynamical map over the time interval with the aim of simulation such that we can apply Φ_{t_{k+1},t_k} for $\{t_k\}_{k=1}^M$, where, for simplicity, we have $t_{k+1} = t_k + \Delta t$, with $\Delta t = T/M$. The corresponding sequence of states we seek to learn are the evolved quantum states at each time step,

$$\hat{\sigma}(t_k) = \Phi_{t_k,0}(\hat{\sigma}_0) = \Phi_{t_k,t_{k-1}} \circ \Phi_{t_{k-1},t_{k-2}} \circ \ldots \circ \Phi_{t_1,0}(\hat{\sigma}_0). \tag{3.48}$$

The naive approach would be to simply identify each target state as $\hat{\sigma}(\boldsymbol{\Lambda}(\tau_k)) = \hat{\sigma}(t_k)$ and formulate the problem as a specific instantiation of meta-QMHL (3.47). However, we note that to construct each $\hat{\sigma}(t_k)$, the quantum circuit depth grows linearly with k. The quantum variational recursive time evolution ansatz (QVARTZ) aims to circumvent this scaling by recursively learning our QHBM representations. Given the optimal QHBM at the previous time step $\hat{\rho}_{\boldsymbol{\Omega}^*(t_{k-1})}$, we apply the single channel for the current time step $\Phi_{t_k,t_{k-1}}$ and learn the current model $\hat{\rho}_{\boldsymbol{\Omega}(t_k)}$ against the resulting evolved state, which serves as approximation of the true evolved state,

$$\hat{\sigma}(t_k) \approx \Phi_{t_k,t_{k-1}}(\hat{\rho}_{\boldsymbol{\Omega}^*(t_{k-1})}). \tag{3.49}$$

Formally, we set $\hat{\sigma}(\boldsymbol{\Lambda}(\tau_k)) = \Phi_{t_k,t_{k-1}}(\hat{\rho}_{\boldsymbol{\Omega}^*(\tau_{k-1})})$ in the meta-QMHL objective (3.47). We may intuitively view this approach as checkpointing the quantum dynamics of a system in the classical parameters of a QHBM. As a result, our quantum circuit depth requirements are now constant with respect k for we no longer need to repeatedly propagate our initial state through a series of channels at each step and can instead initialize the evolution from our latest QHBM representation.

Chapter 4

Experiments

4.1 Transverse-Field Ising Model

For our experiments, we choose to study the transverse-field Ising model (TFIM), which describes a quantum system of spins on a lattice featuring nearest neighbor ferromagnetic interactions along the z axis in addition to an external magnetic field directed along the transverse x axis. The Hamiltonian for this system is given by

$$\hat{H}_{\mathrm{TFIM}} = -\sum_{<i,j>} \hat{Z}_i \hat{Z}_j - h \sum_i \hat{X}_i, \tag{4.1}$$

where the first summation is performed over pairs of nearest neighboring lattice sites and h denotes the relative strength of the transverse field as compared to the nearest neighbor interaction.

4.2 Model Architecture

We utilize a particular QHBM architecture that in principle can universally approximate any density operator in the limit of many variational parameters, which we accordingly refer to as the basic universal density operator ansatz (BUDA). For the choice of energy function of the classical energy-based model, we consider a K^{th}-order binary energy function (KOBE) of the form

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{b} \in \mathcal{B}_K} \theta_{\boldsymbol{b}} (-1)^{\|\boldsymbol{b} \cdot \boldsymbol{x}\|} \tag{4.2}$$

with a corresponding latent modular Hamiltonian given by

$$\hat{K}_{\boldsymbol{\theta}} = \sum_{\boldsymbol{b} \in \mathcal{B}_K} \theta_{\boldsymbol{b}} \hat{\boldsymbol{Z}}^{\boldsymbol{b}}, \tag{4.3}$$

where $\|\boldsymbol{b}\| \equiv \sum_{j=1}^{n} |b_j|$ is the Hamming norm, $\mathcal{B}_K = \{\boldsymbol{b} \in \mathbb{Z}_2^n | \|\boldsymbol{b}\| \leq K\}$ is a Hamming ball centered at the origin of radius K, and $\hat{\boldsymbol{Z}}^{\boldsymbol{b}} \equiv \bigotimes_{j=1}^{n} \hat{Z}_j^{b_j}$ is a convenient notation for Pauli

operators. As for the unitary quantum circuit, we employ the standard hardware-efficient ansatz, which can generally be expressed as a product of layers of unitaries,

$$\hat{U}_{\boldsymbol{\phi}} = \prod_{\ell=1}^{L} \hat{V}^{\ell} \hat{U}^{\ell}_{\boldsymbol{\phi}^{\ell}}, \tag{4.4}$$

where each layer consists of a nonparametric unitary of two-qubits entangling gates $\hat{V}^{\ell}$ in addition to a parameterized unitary of single-qubit Pauli rotations,

$$\hat{U}^{\ell}_{\boldsymbol{\phi}^{\ell}} = \bigotimes_{j=1}^{n} e^{-i\boldsymbol{\phi}^{\ell}_{j}\hat{P}^{\ell}_{j}}. \tag{4.5}$$

4.3 Quantum Metric-Aware Descent

We evaluate the proposed metric-aware gradient descent algorithms and compare their performance to other standard optimization techniques on the particular task of simulating the thermal state of the transverse-field Ising model Hamiltonian $\hat{H}_{\text{TFIM}}$ 4.1 via VQT. For purposes of simplicity, we consider a one-dimensional chain of 4 qubits with a transverse field strength of $h = 1$ and an inverse temperature of $\beta = 1$. We variationally optimize the parameters $\boldsymbol{\Omega}$ of a QHBM to minimize the VQT loss 2.6 using vanilla gradient descent, Adam, and quantum-probabilistic natural gradient descent (QPNGD). In practical consideration of the trainability of our model, we avoid utilizing an overly parameterized BUDA architecture, electing instead to use a second-order KOBE function (4.2), which is equivalent to a Boltzmann machine, and a hardware-efficient ansatz (4.4), (4.5) with 2 layers, each consisting of Pauli X and Z rotations on every qubit and controlled-Z gates between pairs of nearest neighboring qubits. All optimization algorithms are run for a total of 10 different random intializations of the variational parameters over 1000 training iterations with a learning rate of 0.01. Any classical or quantum expectation values are estimated using 10^5 samples. In the particular case of QPNGD, we employ a regularization of $\epsilon = 0.01$ to ensure that the transformed sample-based approximation of the information matrix $\mathcal{I}(\boldsymbol{\Omega}_j) + \epsilon\mathbb{1}$ is positive definite. We further choose to directly solve the corresponding linear system rather than explicitly computing the inverse in the interest of numerical stability and efficiency.

The results of our experiments are show in Figure 4.1. Vanilla gradient descent converges on the highest value of the loss function out of all methods as it has a static learning rate and thus no mechanism by which to overcome so-called barren plateaus, characteristic regions in the loss landspaces of variational quantum algorithms with vanishingly small gradients. In comparison, Adam, which is a first-order adaptive method that utilizes an running average of the gradients normalized by a running average of their corresponding magnitudes, eventually reaches a lower value of the loss function, though it appears to struggle on multiple barren plateaus encountered in its optimization path. QPNGD attains a slightly lower value of the loss function, however, it manages to do so in a significantly fewer number of iterations,

given its ability to dynamically tune the step size in accordance with the particular local curvature of the loss landscape.

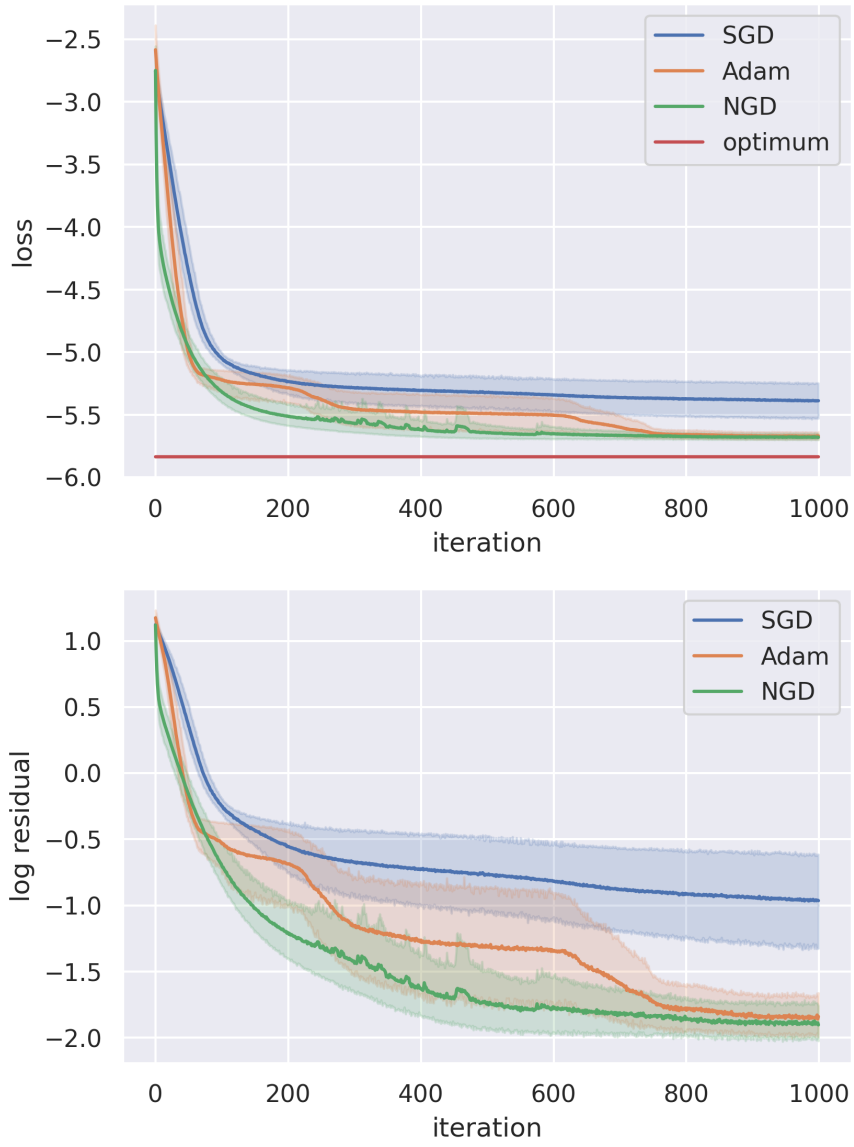

Figure 4.1: Top: the value of the VQT loss as a function of the number of training iterations for the different optimization methods of stochastic gradient descent (SGD), Adam, and natural gradient descent (NGD) compared against the optimum of the loss. Bottom: the logarithm of the residual of the loss for each of the optimization methods.

Bibliography

[1] Shun-Ichi Amari. "Natural gradient works efficiently in learning". In: *Neural computation* 10.2 (1998), pp. 251–276.

[2] Shun-ichi Amari. *Information geometry and its applications*. Vol. 194. Springer, 2016.

[3] Frank Arute et al. "Quantum supremacy using a programmable superconducting processor". In: *Nature* 574.7779 (2019), pp. 505–510.

[4] Koenraad MR Audenaert and Jens Eisert. "Continuity bounds on the quantum relative entropy". In: *Journal of mathematical physics* 46.10 (2005), p. 102104.

[5] Michael Broughton et al. "Tensorflow quantum: A software framework for quantum machine learning". In: *arXiv preprint arXiv:2003.02989* (2020).

[6] Marco Cerezo and Patrick J Coles. "Impact of barren plateaus on the hessian and higher order derivatives". In: *arXiv preprint arXiv:2008.07454* (2020).

[7] Peter Dayan et al. "The helmholtz machine". In: *Neural computation* 7.5 (1995), pp. 889–904.

[8] Richard P Feynman. "Simulating physics with computers". In: *Int. J. Theor. Phys* 21.6/7 (1982).

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[10] Matheus R Grasselli and Raymond F Streater. "On the uniqueness of the Chentsov metric in quantum information geometry". In: *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 4.02 (2001), pp. 173–182.

[11] Hiroshi Hasegawa. "Exponential and mixture families in quantum statistics: Dual structure and unbiased parameter estimation". In: *Reports on Mathematical Physics* 39.1 (1997), pp. 49–68.

[12] Geoffrey E Hinton. "Training products of experts by minimizing contrastive divergence". In: *Neural computation* 14 (2002), pp. 1771–1800.

[13] Ryszard Horodecki et al. "Quantum entanglement". In: *Reviews of modern physics* 81.2 (2009), p. 865.

[14] Bálint Koczor and Simon C Benjamin. "Quantum natural gradient generalised to non-unitary circuits". In: *arXiv preprint arXiv:1912.08660* (2019).

[15] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[16] Tengyuan Liang et al. "Fisher-rao metric, geometry, and complexity of neural networks". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 888–896.

[17] S Lloyd. "Universal Quantum Simulators". en. In: *Science* 273.5278 (Aug. 1996), pp. 1073–1078.

[18] James Martens and Roger Grosse. "Optimizing neural networks with kronecker-factored approximate curvature". In: *International conference on machine learning*. PMLR. 2015, pp. 2408–2417.

[19] Jarrod R McClean et al. "The theory of variational hybrid quantum-classical algorithms". In: *New Journal of Physics* 18.2 (2016), p. 023023.

[20] Alberto Peruzzo et al. "A variational eigenvalue solver on a photonic quantum processor". In: *Nature communications* 5 (2014), p. 4213.

[21] Denes Petz. "Bregman divergence as relative operator entropy". In: *Acta Mathematica Hungarica* 116.1 (2007), pp. 127–131.

[22] Dénes Petz and Catalin Ghinea. "Introduction to quantum Fisher information". In: *Quantum probability and related topics*. World Scientific, 2011, pp. 261–281.

[23] John Preskill. "Quantum Computing in the NISQ era and beyond". In: *Quantum* 2 (2018), p. 79.

[24] Garvesh Raskutti and Sayan Mukherjee. "The information geometry of mirror descent". In: *IEEE Transactions on Information Theory* 61.3 (2015), pp. 1451–1457.

[25] James Stokes et al. "Quantum natural gradient". In: *Quantum* 4 (2020), p. 269.

[26] Barnaby van Straaten and Bálint Koczor. "Measurement cost of metric-aware variational quantum algorithms". In: *arXiv preprint arXiv:2005.05172* (2020).

[27] Geza Toth. "Lower bounds on the quantum Fisher information based on the variance and various types of entropies". In: *arXiv preprint arXiv:1701.07461* (2017).

[28] Hisaharu Umegaki. "Conditional expectation in an operator algebra, IV (entropy and information)". In: *Kodai Mathematical Seminar Reports*. Vol. 14. Department of Mathematics, Tokyo Institute of Technology. 1962, pp. 59–85.

[29] Guillaume Verdon et al. "Quantum Hamiltonian-Based Models and the Variational Quantum Thermalizer Algorithm". In: *arXiv preprint arXiv:1910.02071* (2019).

[30] Mark M. Wilde. *Quantum Information Theory*. Second Edition. Cambridge University Press, 2017. ISBN: 9781107176164.

Appendix

A.1 Bogoliubov-Kubo-Mori Information Matrix

We cover how to obtain the Bogoliubov-Kubo-Mori (BKM) metric tensor of the parameters in terms of the QHBM parameterization. Specifically, we provide analytical expressions for sampling-based techniques to obtain unbiased estimates of the matrix elements. We split up our calculation into three types of blocks of this matrix; the cases where the derivatives are both of $\boldsymbol{\theta}$ parameters, the cases where they are both $\boldsymbol{\phi}$ parameters, and the cases where they are a mixture of both types of parameters. The fact that we can computed analytic expressions for the metric tensor for which we can sample the values using a mixture of the quantum and classical computers is unique to the QHBM class of models.

In particular, when resolving the metric to a basis we may use that of the Ω_j tangent vectors so as to assume the parameter space dynamics induced by the QHBM parameterization [27],

$$[\mathcal{I}^{\text{BKM}}(\boldsymbol{\Omega})]_{j,k} = \int_0^\infty \text{tr}\big[(\partial_{\Omega_j}\hat{\rho}_{\boldsymbol{\Omega}})(\hat{\rho}_{\boldsymbol{\Omega}} + s\mathbb{1})^{-1}(\partial_{\Omega_k}\hat{\rho}_{\boldsymbol{\Omega}})(\hat{\rho}_{\boldsymbol{\Omega}} + s\mathbb{1})^{-1}\big]ds \tag{6}$$

$$= \text{tr}\big[(\partial_{\Omega_j}\hat{\rho}_{\boldsymbol{\Omega}})(\partial_{\Omega_k}\log\rho_{\boldsymbol{\Omega}})\big]. \tag{7}$$

We have termed this the Bogolubov-Kubo-Mori (BKM) information matrix.

Probabilistic Block

We first compute the BKM logarithmic derivative,

$$\partial_{\theta_k}\log\rho_{\boldsymbol{\Omega}} = -\partial_{\theta_k}(\hat{K}_{\boldsymbol{\Omega}} + \mathbb{1}\log Z_{\boldsymbol{\theta}}) \tag{8}$$

$$= -U_{\boldsymbol{\phi}}(\partial_{\theta_k}K_{\boldsymbol{\theta}})U_{\boldsymbol{\phi}}^\dagger + \frac{1}{Z_{\boldsymbol{\theta}}}\text{tr}\big[(\partial_{\theta_k}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}}\big] \tag{9}$$

and tangent vector,

$$\partial_{\theta_k}\rho_{\mathbf{\Omega}} = U_{\boldsymbol{\phi}}\left(\partial_{\theta_k}\frac{e^{-K_{\boldsymbol{\theta}}}}{Z_{\boldsymbol{\theta}}}\right)U_{\boldsymbol{\phi}}^{\dagger} \tag{10}$$

$$= U_{\boldsymbol{\phi}}\frac{-Z_{\boldsymbol{\theta}}(\partial_{\theta_k}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}} + e^{-K_{\boldsymbol{\theta}}}\operatorname{tr}\left[(\partial_{\theta_k}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}}\right]}{Z_{\boldsymbol{\theta}}^2}U_{\boldsymbol{\phi}}^{\dagger} \tag{11}$$

$$= U_{\boldsymbol{\phi}}\left(-(\partial_{\theta_k}K_{\boldsymbol{\theta}})\rho_{\boldsymbol{\theta}} + \rho_{\boldsymbol{\theta}}\operatorname{tr}[(\partial_{\theta_k}K_{\boldsymbol{\theta}})\rho_{\boldsymbol{\theta}}]\right)U_{\boldsymbol{\phi}}^{\dagger} \tag{12}$$

Therefore,

$$[\mathcal{I}^{\mathrm{BKM}}(\mathbf{\Omega})]_{\theta_j,\theta_k} = \operatorname{tr}\left[\frac{(\partial_{\theta_j}K_{\boldsymbol{\theta}})(\partial_{\theta_k}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}}}{Z_{\boldsymbol{\theta}}}\right] - \frac{\operatorname{tr}\left[(\partial_{\theta_j}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}}\right]\operatorname{tr}\left[(\partial_{\theta_k}K_{\boldsymbol{\theta}})e^{-K_{\boldsymbol{\theta}}}\right]}{Z_{\boldsymbol{\theta}}^2} \tag{13}$$

$$= \sum_x p_{\boldsymbol{\theta}}(x)\partial_{\theta_j}E_{\boldsymbol{\theta}}(x)\partial_{\theta_k}E_{\boldsymbol{\theta}}(x) - \sum_x p_{\boldsymbol{\theta}}(x)\partial_{\theta_j}E_{\boldsymbol{\theta}}(x)\sum_y p_{\boldsymbol{\theta}}(y)\partial_{\theta_k}E_{\boldsymbol{\theta}}(y) \tag{14}$$

$$= \mathbb{E}_{x\sim p_{\boldsymbol{\theta}}(x)}\left[\partial_{\theta_j}E_{\boldsymbol{\theta}}(x)\partial_{\theta_k}E_{\boldsymbol{\theta}}(x)\right] - \mathbb{E}_{x\sim p_{\boldsymbol{\theta}}(x)}\left[\partial_{\theta_j}E_{\boldsymbol{\theta}}(x)\right]\mathbb{E}_{y\sim p_{\boldsymbol{\theta}}(y)}\left[\partial_{\theta_k}E_{\boldsymbol{\theta}}(y)\right] \tag{15}$$

The result reads as the covariance matrix of the gradient vector of the energy function subject to the sampled EBM distribution. Note that this quantity does not require a quantum computer to be evaluated.

Quantum Block

For the BKM metric tensor elements which only depend on the gradients with respect to the QNN parameters, we can use an intuitive double parameter shift rule. A gradient technique for QNNs was recently pointed out in [6]; here we can apply it to the gradients of the QHBM QNN parameters. For a hardware efficient ansatz, we have the parameter shift rules,

$$\partial_{\phi_k}\hat{K}_{\boldsymbol{\theta}\boldsymbol{\phi}} = \hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^k)} - \hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^k)} \tag{16}$$

$$\partial_{\phi_k}\hat{\rho}_{\boldsymbol{\theta}\boldsymbol{\phi}} = \hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^k)} - \hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^k)} \tag{17}$$

$$\begin{aligned}[\mathcal{I}^{\mathrm{BKM}}(\mathbf{\Omega})]_{\phi_j,\phi_k} = {} & \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^j)}\hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^k)}\right] + \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^j)}\hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^k)}\right] \\ & - \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^j)}\hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^k)}\right] - \operatorname{tr}\left[\hat{\rho}_{\boldsymbol{\theta}(\boldsymbol{\phi}-\mathbf{\Delta}^j)}\hat{K}_{\boldsymbol{\theta}(\boldsymbol{\phi}+\mathbf{\Delta}^k)}\right]\end{aligned} \tag{18}$$

with $\mathbf{\Delta}^j = \frac{\pi}{4}\hat{\boldsymbol{e}}_{\boldsymbol{j}}$ where standard basis vector has entries $(\hat{\boldsymbol{e}}_{\boldsymbol{j}})_k = \delta_{j,k}$.

Cross Block

Finally, let us compute the terms of the BKM metric tensor which include the coupling of QNN and EBM parameters,

$$[\mathcal{I}^{\mathrm{BKM}}(\boldsymbol{\Omega})]_{\phi_j,\theta_k} = -\operatorname{tr}\Big[(\partial_{\theta_k} K_{\boldsymbol{\theta}}) U_{\boldsymbol{\phi}}^{\dagger} U_{\boldsymbol{\phi}+\boldsymbol{\Delta}^j} \hat{\rho}_{\boldsymbol{\theta}} U_{\boldsymbol{\phi}+\boldsymbol{\Delta}^j}^{\dagger} U_{\boldsymbol{\phi}}\Big] + \operatorname{tr}\Big[(\partial_{\theta_k} K_{\boldsymbol{\theta}}) U_{\boldsymbol{\phi}}^{\dagger} U_{\boldsymbol{\phi}-\boldsymbol{\Delta}^j} \hat{\rho}_{\boldsymbol{\theta}} U_{\boldsymbol{\phi}-\boldsymbol{\Delta}^j}^{\dagger} U_{\boldsymbol{\phi}}\Big], \tag{19}$$

where we have essentially combined eqs. (8) and (16).