Quantum Earth Mover's Distance: A New Approach to Learning Quantum Data

Bobak Toussi Kiani[1,2], Giacomo De Palma[2], Milad Marvian[3], Zi-Wen Liu[4], and Seth Lloyd[2,5]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA
[4]Perimeter Institute for Theoretical Physics, Waterloo, Canada
[5]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Quantifying how far the output of a learning algorithm is from its target is an essential task in machine learning. However, in quantum settings, the loss landscapes of commonly used distance metrics often produce undesirable outcomes such as poor local minima and exponentially decaying gradients. As a new approach, we consider here the quantum earth mover's (EM) or Wasserstein-1 distance, recently proposed in [De Palma *et al.*, arXiv:2009.04469] as a quantum analog to the classical EM distance. We show that the quantum EM distance possesses unique properties, not found in other commonly used quantum distance metrics, that make quantum learning more stable and efficient. We propose a quantum Wasserstein generative adversarial network (qWGAN) which takes advantage of the quantum EM distance and provides an efficient means of performing learning on quantum data. Our qWGAN requires resources polynomial in the number of qubits, and our numerical experiments demonstrate that it is capable of learning a diverse set of quantum data.

1 Introduction

A fundamental task in quantum machine learning is designing efficient algorithms for learning quantum states [1–10], transformations [11–16], and classical data stored as or generated by quantum states [17–19]. In the general setup, one is given a target quantum object, say a target quantum state, and aims to generate or approximate that target object by efficiently

learning parameters in a quantum circuit. For example, quantum generative adversarial networks (qGAN) are parameterized sets of quantum circuits and quantum operators designed to learn target states or transformations via optimization over the parameters of a quantum generator and discriminator [7, 9].

A crucial component of a quantum machine learning algorithm is a distance metric which determines how close a generated object is to its target. This choice of distance metric is important not only as a measure of performance but also as a means for optimization. For example, certain metrics lead to efficient algorithms for calculating gradients with respect to that metric, allowing algorithms to perform optimization via gradient based optimizers (*e.g.,* gradient descent). Naturally, for learning pure states, a common choice for the distance metric is a function of the inner product between quantum states. Similarly, for learning density matrices, researchers commonly choose distance metrics which simplify to a function of the inner product when measuring the distance between pure states.

Previous approaches to learning quantum data have typically suffered from the presence of vanishing gradients [20–22] and poor local minima [23, 24] in the loss landscape induced by the choice of distance metric. Intuitively, these "barren plateaus" and traps arise due to the fact that random quantum states have inner product that diminishes exponentially with the number of qubits. Our approach helps surmount these challenges by formulating an algorithm which provides an efficient means for learning pure and mixed states using the recently proposed quantum earth mover's (EM) distance, also known as the quantum Wasserstein distance of order 1 [25]. As we will demonstrate, the quantum EM distance is a natural distance metric for optimization over local operations and avoids common pitfalls faced by other distance metrics which reduce to functions of the inner product. This is in agreement with results in classical machine learning, where algorithms employing the earth mover's distance are often more stable and avoid issues with vanishing or exploding gradients [26–30] (see Appendix A for a more complete discussion of the literature and Appendix B for a presentation of the classical EM distance). Intuitively, the quantum EM distance can be interpreted as a continuous version of a quantum "hamming distance", which allows local gates to optimize over the few qubits on which they act instead of over some global distance metric which often decays exponentially in the number of qubits.

We proceed as follows. First, we analyze the properties of different quantum distance metrics and the loss landscapes they produce in quantum machine learning settings. Here, we show that the quantum EM distance has unique advantages over other common distance metrics. Then, to apply the quantum EM distance in a machine learning algorithm, we devise an efficient means to approximately calculate the quantum EM distance. This leads to our development of a quantum Wasserstein generative adversarial network (qWGAN) which is a quantum analog to the classical Wasserstein generative adversarial network [26] (see also [9]). Importantly, like its classical analog, our qWGAN employs an earth mover's distance in its cost function. Numerical results show that our qWGAN is efficient at learning quantum data in various settings. Finally, we discuss near term applications of our qWGAN for both classical and quantum problems.

2 Quantum Distance Metrics

To approximate or reconstruct a target probability distribution with a machine learning algorithm, the choice of distance metric, measuring how well the approximating distribution matches the target distribution, is crucial to the performance of the algorithm. Classically, generative adversarial networks (GAN) provide a neural network approach for learning a target probability distribution and generating new samples from the approximate distribution [26, 31]. The choice of loss metric for a GAN is a distance or divergence metric which is minimized when the target and generated distributions coincide.

In the quantum setting, distance metrics between states or density matrices are employed in the implementation of quantum generative adversarial networks (qGAN) [7, 9, 32]. As in the classical setting, the choice of distance metric is crucial to the runtime and performance of the quantum machine learning algorithm. Here, we consider common distance metrics and show that the quantum earth mover's (EM) distance recently defined in [25] possesses desirable properties that are not found in the other metrics.

For a brief overview of the notation used in quantum mechanics, we refer the reader to Appendix C. Let $\rho, \sigma \in \mathbb{C}^{N\times N}$ be the density matrices corresponding to two quantum states, *e.g.,* ρ can be the quantum state generated by a GAN and σ is the target state. Until now, common distance metrics employed to train quantum GANs have been unitarily invariant, *i.e.*, invariant with respect to the conjugation of both quantum states with the same unitary matrix and reducing to a function of the inner product for pure states (*i.e.,* orthogonal projectors with rank one). Commonly used distance metrics in prior works include:

- **Trace Distance:** The simplest and most common choice (*e.g.,* see [1, 32]) is the trace distance:
$$D_1(\rho, \sigma) = \frac{1}{2} \|\rho - \sigma\|_1 \ , \tag{1}$$
where $\|\cdot\|_1$ denotes the trace norm, *i.e.*, the sum of the singular values.

- **Quantum Fidelity:** Another common choice (*e.g.,* see [10]) is the maximum absolute value squared of the inner product between purifications of ρ and σ:
$$F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1^2 \ . \tag{2}$$
$F(\rho, \sigma)$ is often modified to $\arccos\sqrt{F(\rho, \sigma)}$ to construct a proper distance metric.

- **Quantum Wasserstein Semimetric:** Introduced in [9] as a quantum generalization of the Wasserstein distance, this distance, denoted $qW(\rho, \sigma)$, is calculated by forming a coupling between quantum states ρ and σ in $\mathbb{C}^{N\times N}$. The coupling is a quantum state in $\left(\mathbb{C}^{N\times N}\right)^{\otimes 2}$ whose marginal states are equal to ρ and σ, respectively. The quantum Wasserstein semimetric is the minimum of the expectation value of the projector onto the symmetric subspace of $\left(\mathbb{C}^{N}\right)^{\otimes 2}$. qW does not satisfy the triangle inequality, hence the name semimetric. Importantly, qW is unitarily invariant, and for pure states, it reduces to a function of their inner product: for any $|u\rangle$, $|v\rangle$ unit vectors in $\mathbb{C}^N$, $qW(|u\rangle\langle u|, |v\rangle\langle v|) = (1 - |\langle v|u\rangle|^2)/2$. Further details can be found in Appendix D.

2.1 The quantum EM distance

In this paper we consider the case of n qubits, where $N = 2^n$, and employ the quantum generalization of the Wasserstein distance of order 1 to the states of n qubits recently proposed in [25] and also known as the earth mover's (EM) distance. We adopt the latter terminology as it is more prevalent in the machine learning community, hereby denoting the quantum EM distance with D_{EM}. Unlike all the previously employed distances, the quantum EM distance is not unitarily invariant. We will show that, similar to its classical counterpart [26], D_{EM} possesses several properties that are desirable when learning quantum data.

The quantum EM distance of [25] is based on the notion of neighboring states. Two quantum states of n qubits are neighboring if they differ in only one qubit, *i.e.*, if they coincide after one qubit is discarded. The quantum EM distance is the distance that is induced by the maximum norm that assigns distance at most one to any couple of neighboring states. We denote with $\|\cdot\|_{EM}$ the corresponding norm, whose analytical expression can be found in Appendix E. This definition enforces the continuity of the distance with respect to local operations, *i.e.*, any quantum operation acting on a single qubit can displace a state by at most one unit with respect to the quantum EM distance. Indeed, for the quantum states of the computational basis the quantum EM distance recovers the classical Hamming distance, *i.e.*, for any two strings of n bits x and y we have $D_{EM}(|x\rangle\langle x|, |y\rangle\langle y|) = h(x, y)$. More generally, for quantum states diagonal in the computational basis, the quantum EM distance recovers the classical EM distance. The quantum EM distance admits a dual formulation [25], based on the quantum generalization of the Lipschitz constant, which is more suitable for implementation of quantum GANs.

We denote with $\mathcal{O}_n$ the set of n-qubit observables, *i.e.*, the set of the $2^n \times 2^n$ Hermitian matrices. The quantum Lipschitz constant of the observable $H \in \mathcal{O}_n$ is

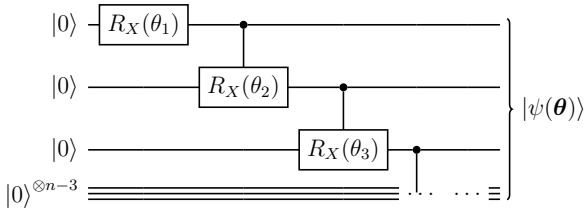
$$\|H\|_L = 2 \max_{i=1,\ldots,n} \min \left(\|H - H_{\bar{i}}\|_\infty : H_{\bar{i}} \in \mathcal{O}_n \text{ does not act on the } i\text{-th qubit} \right) . \tag{3}$$

The quantum Lipschitz constant defined above is a generalization of the Lipschitz constant for the functions on strings of n bits, and coincides with the classical Lipschitz constant for the observables that are diagonal in the computational basis [25]. The quantum EM distance between the quantum states ρ and σ is equal to the maximum difference between the expectation values on ρ and σ of a quantum observable with Lipschitz constant at most one:

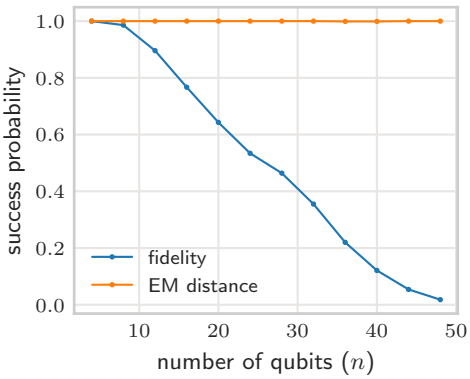
$$D_{EM}(\rho, \sigma) = \max \left(\mathrm{Tr} \left[(\rho - \sigma) H \right] : H \in \mathcal{O}_n, \ \|H\|_L \leq 1 \right) . \tag{4}$$

When the quantum EM distance plays the role of a cost function in a machine learning algorithm, it can be considered as an energy associated to the parameter configuration. For this reason, we may refer to the observables H in (4) as Hamiltonians.

To show why D_{EM} possesses desirable properties, we first consider the case where both the target σ and the generated ρ are pure states in a simple toy model. Here, as we will show, undesirable critical points are clearly present and endemic to the loss landscapes for metrics which are a function of the inner product between two pure states. In contrast, the quantum EM distance D_{EM} avoids these undesirable critical points. Finally, we generalize the findings of this toy model to a larger class of quantum machine learning settings.



(a)



(b)

Figure 1: (a) Simple quantum circuit which can generate the GHZ state when parameter values are appropriately chosen. (b) In learning settings, this circuit presents challenges for loss metrics that are a function of the inner product between target and initial states. Simulations show that, with a loss function such as fidelity (blue line), gradient based optimizers eventually fail to find global optimum in virtually all instances when circuit contains many qubits, instead converging to the state $|0_n\rangle$. In contrast, the use of the quantum EM distance (orange line) results in convergence to the global optimum in virtually all instances tested. Here, optimization is performed for up to 10^5 steps using the Adam optimizer (simulations with EM distance generally achieved convergence within about 1000 steps). Experiments are repeated 1000 times for each n to estimate success probability. See Appendix F for full details of experiments.

2.2 A simple toy model

In this section, we consider an intuitive example which shows the advantages of using the quantum EM distance when the learning is performed over local quantum gates. Namely, we show that the commonly used distance metrics which are a function of the inner product between states feature two key issues in learning via local gates. First, the inner product between a generated and target state fails to show improvement when gates are optimized one by one or layer-wise. Second, when parameters are initialized randomly, gradients in this example decay exponentially when the distance metric is a function of the inner product. The quantum EM distance avoids both of these drawbacks and allows for efficient learning in this scenario.

In this toy model, the task at hand is to learn the correct values of parameters in the circuit in Figure 1a to generate the GHZ state of n qubits $|GHZ_n\rangle = (|0_n\rangle + |1_n\rangle)/\sqrt{2}$. This circuit consists of a parameterized Pauli X rotation on the first qubit and controlled parameterized Pauli X rotations on later qubits (see Appendix C for description of Pauli operators). When n is a multiple of 4, setting θ_1 equal to $\pi/2$ and all other parameters equal to π will construct the target GHZ state.

Consider the case where one aims to maximize the fidelity F between the generated state

$|\psi(\boldsymbol{\theta})\rangle$ and the GHZ state $|GHZ_n\rangle$. Given our circuit, F takes a simple form:

$$F = |\langle GHZ_n|\psi(\boldsymbol{\theta})\rangle|^2 = \left(\frac{\cos(\theta_1)}{\sqrt{2}} + \frac{\prod_{i=1}^{n}\sin(\theta_i)}{\sqrt{2}}\right)^2. \tag{5}$$

The first problem associated with learning via F is that the loss landscape has many local minima associated with the target state $|0_n\rangle$. Note that fixing any $\theta_i = 0$ will force a learning algorithm (*e.g.,* gradient descent) to optimize the $\cos(\theta_1)$ term, converging to the state $|0_n\rangle$. In other words, if any algorithm aims to optimize the gates in a layer-wise fashion (*e.g.,* optimizing $\theta_1, \theta_2, \ldots$ in order as in [33]), that algorithm will get stuck in the local optimum at $|0_n\rangle$.

Of course, in practice, parameters $\boldsymbol{\theta}$ are typically initialized randomly so it is unlikely that any $\theta_i = 0$; however, even here, we have the second issue that gradients with respect to $\theta_2, \theta_3, \ldots, \theta_n$ all decay exponentially with n since the product of sine functions with random parameter values decays exponentially to zero.

$$\frac{dF}{d\theta_i} = \begin{cases} -\cos\theta_1\sin\theta_1 + \cos 2\theta_1 \prod_{k=2}^{n}\sin\theta_k & i = 1 \\ \quad + \sin\theta_1\cos\theta_1\left(\prod_{k=2}^{n}\sin\theta_k\right)^2 & \\ \cos\theta_i \prod_{k\neq i}\sin(\theta_k)\left[\cos(\theta_1) + \prod_{k=1}^{n}\sin(\theta_k)\right] & i > 1 \end{cases}. \tag{6}$$

Notably, $\frac{dF}{d\theta_i} = O(\frac{1}{2^n})$ for $i > 1$, but $\frac{dF}{d\theta_i} = O(1)$ for $i = 1$. For large n, $\frac{dF}{d\theta_1} \approx -\cos\theta_1\sin\theta_1$ indicating that gradient optimizers will converge to a poor local minimum outputting the state $|0_n\rangle$ ($\theta_1 = 0 \mod 2\pi$). In fact, since the loss function F (equation (5)) takes a simple form, gradient descent on the parameters can be efficiently performed classically, and results shown in Figure 1b show that gradient descent converges almost surely to the undesirable local optimum associated with the $|0_n\rangle$ state as more qubits are added (simulations stopped after 100000 steps of optimization).

The challenges described above are encapsulated by the feature of the inner product that, for example, the states $|0000\rangle$, $|1000\rangle$, and even $|1110\rangle$ are equally distant (orthogonal) to the state $|1111\rangle$ – *i.e.,* local updates induce no change in the inner product distance metric. Using a loss function with the quantum EM distance D_{EM} naturally avoids these challenges. Since the quantum EM distance recovers the Hamming distance between two computational basis states, local operations on one and two qubits can reduce D_{EM} as long as those operations reduce the Hamming distance between the target and generated states. For example, unlike the inner product distance metric, $|1000\rangle$ and $|1110\rangle$ are three and one units away respectively from the state $|1111\rangle$ in the quantum EM distance. In our toy model, local gates, even when applied in isolation, result in changes to single qubits which either reduce or increase the quantum EM distance.

If we set $\theta_1 = \pi/2$, $\theta_2 = \ldots = \theta_k = \pi$ and $\theta_{k+1} = \ldots = \theta_n = 0$ in the quantum circuit of Figure 1a, we obtain the quantum state $|\Psi_k\rangle = \left(|0_k\rangle + (-i)^k|1_k\rangle\right)|0_{n-k}\rangle/\sqrt{2}$. The following Proposition 1 shows that the sequence of states $|\Psi_0\rangle = |0_n\rangle, |\Psi_1\rangle, \ldots, |\Psi_n\rangle = |GHZ_n\rangle$ gets closer and closer to the target state $|GHZ_n\rangle$, with a guaranteed improvement every two steps. The proof is in Appendix G.

Proposition 1. *For any $k = 0, \ldots, n$, let $D_k = D_{EM}\left(|\Psi_k\rangle\langle\Psi_k|, |GHZ_n\rangle\langle GHZ_n|\right)$. We have $n/2 \leq D_0 \leq (n+1)/2$, $D_n = 0$, and $(n-k)/2 \leq D_k \leq \left(n-k+\sqrt{2}\right)/2$ for any $k = 1, \ldots, n-1$. In particular, we have $D_{k+2} < D_k$ for any $k = 0, \ldots, n-2$.*

Furthermore, as shown in Figure 1b, optimization using the quantum EM distance is, in virtually all cases, successful at learning the GHZ state. In the simulations for Figure 1b, the quantum EM distance is efficiently estimated using the dual formulation by considering the expectation of the generated state over a subset of $O(n)$ Hermitian operators H_i all with Lipschitz constant equal to one.

$$\begin{aligned}\tilde{D}_{EM} &= \max_{H_i} \left| \langle\psi(\boldsymbol{\theta})| H_i |\psi(\boldsymbol{\theta})\rangle - \langle GHZ_n| H_i |GHZ_n\rangle \right| \\ &\approx D_{EM}\left(|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|, |GHZ_n\rangle\langle GHZ_n|\right),\end{aligned} \tag{7}$$

where $\tilde{D}_{EM}$ is the approximation which lower bounds D_{EM} by taking the maximum over the $O(n)$ operators H_i chosen for optimization of the circuit (see Appendix F for list of operators). Using $\tilde{D}_{EM}$, we can successfully learn and generate the GHZ state regardless of the size of the system. Given the simplified form of our circuit, calculating $\tilde{D}_{EM}$ can be efficiently performed using a classical computer and the methodology is detailed in Appendix F. Interestingly, though the subset of Hamiltonians considered in calculating $\tilde{D}_{EM}$ is significantly less than the total space of Hamiltonians available needed to exactly calculate D_{EM}, the simplified form of $\tilde{D}_{EM}$ still suffices to completely learn the GHZ state. This perhaps surprising fact is one motivation for our qWGAN, discussed later, which uses similar techniques to construct a general algorithm for learning quantum data in more complex settings.

2.3 Generalizing findings of toy model

As our toy model shows, the quantum EM distance confers distinct advantages when optimizing over local parameters. To generalize these findings, we discuss unique properties of the quantum EM distance not necessarily found in other distance metrics. First, the quantum EM distance is super-additive with respect to the tensor product [25, Proposition 3]:

Proposition. *For any two quantum states ρ, σ of n qubits and any $k = 1, \ldots, n-1$,*

$$D_{EM}(\rho, \sigma) \geq D_{EM}(\rho_{1\ldots k}, \sigma_{1\ldots k}) + D_{EM}(\rho_{k+1\ldots n}, \sigma_{k+1\ldots n}), \tag{8}$$

where $\rho_{1\ldots k}$ and $\rho_{k+1\ldots n}$ are the marginal states of ρ over the first k and the last $n-k$ qudits, respectively, and analogously for σ.

Intuitively, the Proposition above implies that operations which reduce the distance between two states over a portion of their qubits will proportionally reduce the total distance over all of the qubits. Note that no unitarily invariant distance can have this property. For example, to learn a target state $|GHZ_2\rangle|1\rangle$, updating the state $|000\rangle$ to $|GHZ_2\rangle|0\rangle$ results

in a significant improvement in the quantum EM distance but, since the updated state is still orthogonal to the target state, no unitarily invariant distance will show any improvement.

A second useful property of the quantum EM distance is that it recovers the classical earth mover's distance for quantum states diagonal in the canonical basis, and in particular, it recovers the classical Hamming distance for the quantum states of the computational basis [25, Proposition 5]:

Proposition. *Let p, q be probability distributions on $\{0,1\}^n$, and let*

$$\rho = \sum_{x\in\{0,1\}^n} p(x)\,|x\rangle\langle x|\,, \qquad \sigma = \sum_{y\in\{0,1\}^n} q(y)\,|y\rangle\langle y|\,. \tag{9}$$

Then, $D_{EM}(\rho,\sigma) = D_{EM}(p,q)$. In particular, the quantum EM distance between vectors of the canonical basis coincides with the Hamming distance: $D_{EM}(|x\rangle\langle x|, |y\rangle\langle y|) = h(x,y)$ for any x, $y \in \{0,1\}^n$.

The above proposition implies that advantages conferred in classical machine learning algorithms when using the classical EM distance directly translate into quantum settings when using the quantum EM distance. Finally, the quantum EM distance is always contained between the trace distance and n times the trace distance [25, Proposition 6]:

Proposition. *For any two quantum states ρ, σ,*

$$D_1(\rho,\sigma) \leq D_{EM}(\rho,\sigma) \leq n\, D_1(\rho,\sigma)\,. \tag{10}$$

In particular, a small quantum EM distance guarantees that the trace distance is also small and vice-versa. Thus, convergence in the quantum EM distance necessarily implies convergence in more conventional quantum distance metrics such as fidelity or trace distance.

3 EM Distance Evaluation

In practice, the quantum EM distance can be evaluated using quantum algorithms for semidefinite programs [34, 35] which run in time polynomial in the dimension of the quantum state and the number of constraints. However, this approach is neither efficient in time nor does it lead to obvious methods for calculating the gradient of the quantum EM distance. Instead, we provide a procedure below to estimate the quantum EM distance between two distributions of quantum states using its dual formulation (4). To avoid cumbersome computation of Lipschitz constants, as in the classical Wasserstein GAN [26], we construct a parameterized family of functions which preserve this quantum Lipschitz constraint upon optimization. Proposition 9 of [25] provides an upper bound to the quantum Lipschitz constant of a Hamiltonian in terms of its local structure. Let

$$H = \sum_{\mathcal{I}\subseteq\{1,\ldots,n\}} H_{\mathcal{I}}\,, \tag{11}$$

where each $H_{\mathcal{I}}$ acts non-trivially only on the qubits in the corresponding set $\mathcal{I}$. Then, the quantum Lipschitz constant of H is bounded as below:

$$\|H\|_L \leq 2 \max_{i=1,\ldots,n} \left\| \sum_{i \in \mathcal{I} \subseteq \{1,\ldots,n\}} H_{\mathcal{I}} \right\|_\infty , \tag{12}$$

where the maximum is taken over the qubits. The notation $i \in \mathcal{I} \subseteq \{1, \ldots, n\}$ indicates that the sum is taken only over the set of operators which act non-trivially (*i.e.,* not the identity) on qubit i. A natural choice for the operators $H_{\mathcal{I}}$ are a subset of the Pauli operators which we explore in our construction of a quantum generative adversarial network next.

4 qWGAN Algorithm

Our quantum Wasserstein generative adversarial net (qWGAN) consists of a discriminator and generator which approximates a target distribution over states ρ_{tar} by "playing" a min-max game. Here, the generator sets its parameters θ outputting a state $G(\theta)$, and the discriminator $H(W)$ is a parameterized sum of Hermitian operators with weights W. In each iteration of optimization, the discriminator first sets its operator weights, outputting a Hamiltonian $H_{\max}$ which is the Hamiltonian maximizing our dual formulation estimate of $D_{EM}(G(\theta), \rho_{\text{tar}})$. Then, a gradient update is performed on the parameters of the generator θ. This iterative process is repeated either until convergence in the generator parameters θ or until a stopping criterion is reached. We detail the forms of the discriminator and generator as well as the steps of the algorithm in this section.

4.1 Form of the discriminator

In an optimal scenario, a discriminator explores the complete set of Hamiltonians which have Lipschitz constant less than or equal to one. However, this ideal case does not lend itself to efficient algorithms, and we instead construct a discriminator which efficiently estimates (lower bounds) the quantum EM distance. The discriminator we choose is a parameterized sum of strings of Pauli operators:

$$H(W) = \sum_{P_1,\ldots,P_n \in \{I,X,Y,Z\}} w_{P_1 \ldots P_n} \, \sigma_{P_1}^{(1)} \otimes \sigma_{P_2}^{(2)} \otimes \ldots \otimes \sigma_{P_n}^{(n)} , \tag{13}$$

where σ_I is the 2×2 identity matrix, σ_X, σ_Y and σ_Z are the Pauli matrices, superscripts specify the qubit on which the corresponding Pauli matrix acts and each $w_{P_1 \ldots P_n}$ is the trainable parameter for the corresponding Pauli string $\sigma_{P_1}^{(1)} \otimes \sigma_{P_2}^{(2)} \otimes \ldots \otimes \sigma_{P_n}^{(n)}$. To simplify notation, we denote the set of all trainable parameters as W.

The Hamiltonian (13) has 4^n parameters and is impractical to train. For this reason, we restrict optimization to operators that contain only terms acting on few qubits. One option

is to choose the set of k-local Pauli operators as the discriminator. In the case where $k = 2$, we have the following construction:

$$H(W) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{P,\, Q \in \{I,X,Y,Z\}} w_{P,\, Q}^{(i,j)}\, \sigma_P^{(i)} \otimes \sigma_Q^{(j)}, \tag{14}$$

where each $w_{P,\, Q}^{(i,j)}$ is the trainable parameter for the corresponding Pauli operator.

For $k \ll n$, there are $O(n^k)$ total terms in the above summation, polynomial in the number of qubits. To optimize the EM distance using this discriminator, we setup a linear program which can be efficiently solved using classical algorithms for linear programming. To simplify notation, we assume all parameters are enumerated in a list $W = \{w_1, w_2, \ldots, w_{|W|}\}$. For each parameter w_i, we let $\mathcal{I}_i$ be equal to the set of qubits which the corresponding Pauli string acts on. Thus, with $|W|$ parameters and n qubits, one maximizes the following linear program:

$$\begin{aligned} &\text{maximize} && \sum_{j=1}^{|W|} c_j w_j \\ &\text{subject to} && \sum_{j:i\in\mathcal{I}_j} |w_j| \leq 1, \qquad i = 1, ..., n \end{aligned} \tag{15}$$

where c_j is the trace of the product between the j-th Pauli string and $G(\theta) - \rho_{\text{tar}}$. *i.e.,* assuming w_j is associated to Pauli string $\sigma_{P_a}^{(a)} \sigma_{P_b}^{(b)} \ldots \sigma_{P_k}^{(k)}$, then $c_j = \text{Tr}\left[(G - \rho_{\text{tar}}) \sigma_{P_a}^{(a)} \sigma_{P_b}^{(b)} \ldots \sigma_{P_k}^{(k)}\right]$. In the above formulation, there exists a constraint for each qubit i limiting the sum of magnitudes of operators acting on that qubit to less than or equal to one.

The linear program in (15) can be transformed into a standard form linear program with n constraints (one for each qubit), which outputs a sparse set of at most n non-zero weights [36] (the number of non-zero variables in linear programs in standard form is at most the number of constraints). Specifically, the linear program will output $n_{\text{active}} \leq n$ operators with non-zero weights, called active operators, constructing a Hamiltonian $H_{\max}$ which is passed onto the generator for optimization:

$$H_{\max} = \sum_{i=0}^{n_{\text{active}}} w_i' H_i', \tag{16}$$

where w_i' and H_i' are the weights and active operators respectively. Since $n_{\text{active}} \leq n$, this will allow for efficient gradient updates on the generator which aims to minimize the expectation of $H_{\max}$.

As we show in Appendix H, restricting the optimization over operators to terms acting on few qubits does not affect the value of the distance, since the optimal unconstrained Hamiltonian for the maximization problem (15) contains only these terms whenever the coefficients c_j associated to single Pauli operators are all $\Omega(1)$.

Optional cycling of operators Over a small number of steps of optimization, changes to the expectations of operators c_j are expected to be very small. Therefore, if the expectation of a given operator in the discriminator is small, it is unlikely that the operator will be chosen as an active operator over the course of optimization. Therefore, one has the option of removing these "bad" operators and including new, random, operators into the set of operators over which the discriminator optimizes. Many choices exist for cycling operators; here, we opt for a simple choice where operators are cycled out when the expectation of an operator is below a threshold equal to $c(\min_i w_i')$ (*i.e.,* minimum taken over all active operators) where $0 < c \leq 1$. When an operator is cycled out, a random Pauli operator is then included in the discriminator's set of operators.

4.2 Form of the generator

In its most general form, a generator is an object or function, that when given an input (potentially a sample from a random variable), outputs a state which approximates or produces a sample drawn from a distribution close to the target distribution. Similar to classical machine learning where neural networks are customized to given settings – *e.g.,* convolutional neural networks optimized for image analysis [37–39] and transformer networks optimized for text analysis [40–42] – the form of the generator in our quantum algorithm can and should be customized to the specific problem setting. Many options exist for constructing a generator including parameterized quantum circuits [43,44] and quantum neural networks [45–48]. The form of the generator determines the space of functions which a generator can access, and ideally this space should overlap with the function of the target object. Given we can only cover a limited class of generators in our analysis, we focus here on a single, though generic, form for the generator, encouraging future research to construct and analyze generators customized to specific applications in quantum machine learning.

In this generic formulation, the generator $G(\theta)$ is a function which maps a starting state ρ_0 to a density matrix ρ representing the distribution over quantum states that one aims to reconstruct. As in [9], our generator is constructed by a set of probabilities and associated parameterized unitaries $\{(p_1, U_1), \ldots, (p_r, U_r)\}$:

$$G(\theta) = \sum_{i=1}^{r} p_i U_i \rho_0 U_i^\dagger, \tag{17}$$

where we use θ to denote the set of all parameters for the generator which includes the probabilities p_i and parameters for each unitary U_i. r is the maximum rank of the output density matrix which can be tuned as a hyperparameter. Later, we consider U_i constructed by parameterized quantum circuits with one and two qubit gates. The choice of these parameterized circuits depends on the nature of the problem (see section 5 for examples).

4.3 qWGAN optimization procedure

The algorithm for the qWGAN detailed in Algorithm 1 iteratively optimizes parameters of the generator and discriminator, consistent with methods used in classical GANs [26].

The following two steps are repeated until convergence in the parameters of the generator θ. First, the parameters w are updated using the linear program (15) to maximize the quantum EM distance D_{EM} in equation (4). Then, a gradient update is performed on the parameters of the generator θ.

Algorithm 1 qWGAN with quantum earth mover's distance

Require: initial discriminator operators: $H_i^{[0]}$ ▷ *e.g.,* set of 2-local Paulis
Require: initialization of generator parameters: $p_i^{[0]}$ and $\theta_i^{[0]}$
Require: hyperparameters for generator optimizer (*e.g.,* learning rate α)
while θ, p have not converged **do** ▷ alternatively, stop after T steps
  ▷ *discriminator optimization:*
  measure operator expectations: $c_i \leftarrow \mathrm{Tr}\left[H_i(G(\theta) - \rho_{\mathrm{tar}})\right]$
  find w_i', H_i' (linear program, equation (15)) ▷ $H_{\max} = \sum_i w_i' H_i'$
  optional: cycle operators
  ▷ *generator optimization:*
  find gradients g_p, g_θ of $\mathrm{Tr}\left[G(\theta) H_{\max}\right]$ ▷ see Appendix I
  perform gradient update on θ and p ▷ *e.g.,* $\theta \leftarrow \theta - \alpha g_\theta$

5 Simulations

Our qWGAN can efficiently learn quantum data of various forms. Here, we apply the qWGAN directly to the toy model of subsection 2.2 and also consider a more general scenario where the qWGAN learns states generated by a mixing circuit previously known to suffer from barren plateaus [20, 22, 49]. In Appendix J, we include results for the qWGAN in two other scenarios: one where the generator is a circuit formed by a quantum alternating operator ansatz (QAOA) [50–52] and one where the qWGAN is tasked with learning mixed states. Details on the structure of the quantum circuits and on how the simulations were performed are provided in Appendix K and Appendix L, respectively.

5.1 Learning the GHZ state

The n-qubit GHZ state is an entangled state which requires a simple circuit of depth n to construct. However, as noted in subsection 2.2, the correct parameters of this circuit are hard to learn when using cost metrics that are a function of the inner product between the generated and target GHZ state. Continuing our analysis, we show that our qWGAN is especially efficient and effective at learning the correct parameters of a circuit to generate the GHZ state.

As shown in Figure 2, our qWGAN efficiently generates the 8 qubit GHZ state – results for circuits of different size are also consistent with this analysis and detailed in Appendix J. The discriminator starts with access to $k = 2$ local Pauli operators, cycling out "bad" operators

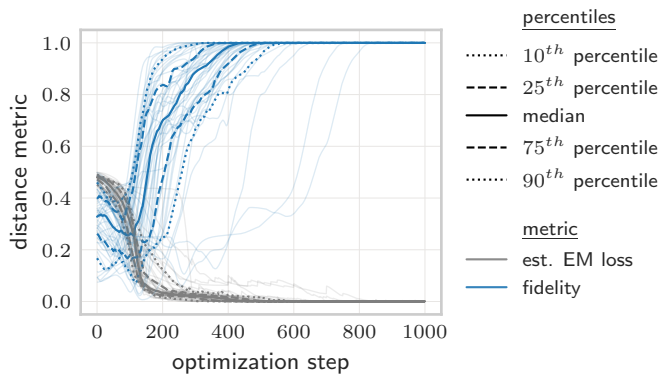


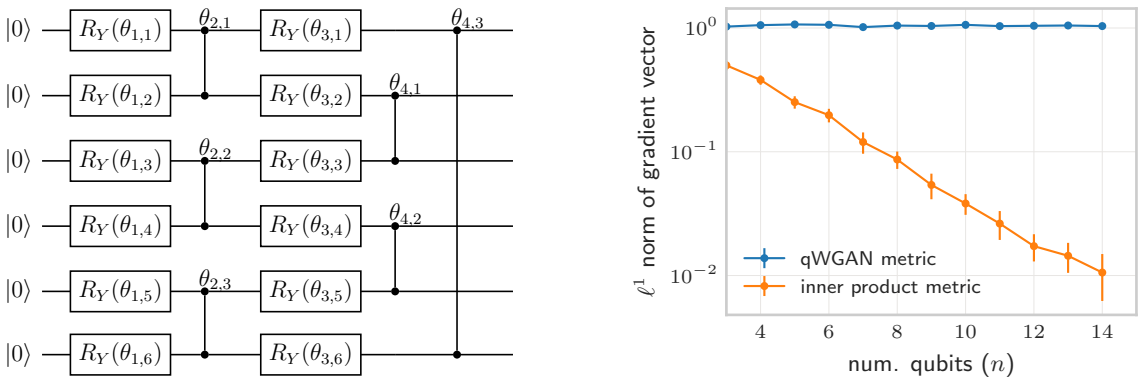
Figure 2: The qWGAN consistently generates the 8 qubit GHZ state. Estimated EM loss (quantum EM distance estimated by active operators) is plotted in blue alongside the fidelity in grey. EM distance is normalized to a maximum of one by dividing by the number of qubits. Percentiles are calculated across 50 simulations. Individual simulations are plotted as transparent lines.

every five steps. Estimated EM loss is also plotted in Figure 2 (normalized by dividing by the number of qubits), which is equal to the quantum EM distance as measured by the active operators in the discriminator (lower bounding the actual quantum EM distance). Jumps in the estimated EM loss can be observed when operators are cycled and later become active, highlighting the importance of randomly cycling operators in these simulations. It is interesting to note that during the early phases of learning, the qWGAN often optimizes the EM distance while temporarily decreasing the fidelity. This learning profile is typically associated with transitions from the state $|0_n\rangle$ to the GHZ state. As our toy model indicated, this transition characterized by a temporary decrease in the fidelity is needed to reach the global optimum.

5.2 Teacher-student learning

To analyze our qWGAN in a more general setting, we consider a "teacher-student" setup where the circuit used to generate the target state and perform learning are both of the form shown in Figure 3a. This circuit is a generic mixing circuit also studied in [20, 22, 49] where barren plateaus in the loss landscape are observed. For our simulations, the target state ϕ_{tar} is generated by a depth 2 circuit (*i.e.* gates shown in Figure 3a repeated twice) with parameters drawn i.i.d. from the standard normal distribution. As a point of comparison, we compare our qWGAN to a quantum GAN equipped with the loss function $F = 1 - |\langle\phi_{\text{tar}}|\phi(\boldsymbol{\theta})\rangle|^2$ which is a function of the inner product between the target and generated state. Figure 3b shows that when a circuit of the same form is used to learn the target state, gradients of F (function of the inner product) decay exponentially with more qubits whereas gradients of the quantum EM loss function remain constant. Note that the exponentially decaying gradients for the inner product loss metrics are observed here for constant depth shallow

circuits. This result further confirms that loss landscapes for the quantum EM distance avoid common pitfalls faced by conventional distance metrics.



(a) (b)

Figure 3: (a) Single layer of mixing circuit consisting of alternating layers of parameterized Pauli Y rotations and parameterized Pauli Z-Z rotations applied to pairwise qubits. Here, the form of the circuit is shown for six qubits. (b) Learning using two layers of mixing circuit (shallow, constant depth) results in exponentially decaying gradients for conventional loss metrics. Gradients of the qWGAN remain constant while gradients with respect to a loss metric as a function of the inner product decay exponentially in the number of qubits. Gradients are calculated at first step of optimization and ℓ^1 norm is divided by n to normalize to the number of parameters in the circuit. Findings are consistent when the average is taken for the ℓ^2 norm or of individual gradient entries as shown in Appendix J. Results are averaged across 100 simulations for each data point.

Furthermore, as shown in Figure 4, the qWGAN successfully learns the states constructed by 8-qubit teacher circuits using student circuits of depth 4. Target states are generated by drawing the parameters of the teacher circuit i.i.d. from the standard normal distribution. Learning is typically achieved within a few hundred steps of optimization. In these simulations, the discriminator for the qWGAN contains all order 2 Pauli operators and no cycling of the operators was performed. Additional simulations for different circuit sizes are shown in Appendix J.

6 Discussion

As interest in quantum machine learning algorithms has flourished, recent research has highlighted the challenges associated with learning using quantum computers. At the root of these challenges are adverse properties of loss landscapes in quantum machine learning settings, perhaps most notably that loss landscapes have poor local minima and exponentially decaying gradients. In this work, we show that the loss landscape induced by the quantum EM distance confers advantages in machine learning settings, especially when optimization

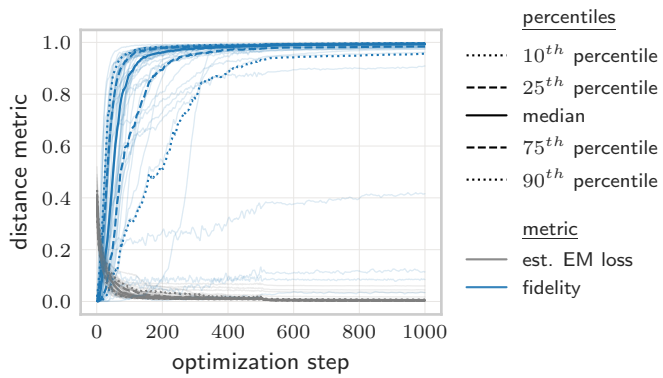


Figure 4: The student circuit is able to approximate well the state generated by the teacher circuit. Here, the target is constructed by randomly setting the parameters of a depth 2 mixing circuit (teacher circuit). The qWGAN, equipped with a depth 4 generator circuit, successfully learns the target state by optimizing over the quantum EM distance. Estimated EM loss (quantum EM distance estimated by active operators) is plotted in blue alongside the fidelity in grey. EM distance is normalized to a maximum of one by dividing by the number of qubits. Percentiles are calculated across 50 simulations. Individual simulations are plotted as transparent lines.

is performed over local gates. Our results provide a new approach to constructing loss landscapes which can avoid common quantum machine learning roadblocks.

For the specific application of learning quantum data, we have proposed a qWGAN which leverages the quantum EM distance to produce an efficient learning algorithm. In accord with its classical counterpart [26], we show that our qWGAN can potentially improve convergence and stability in learning quantum data. Looking beyond the test cases studied here, the qWGAN has many potential applications. In quantum controls, one can use it to search for robust or optimal control parameters [53, 54]. For unsupervised learning, the qWGAN provides a framework and approach to quantum circuit compression, data encoding, and sampling [55–57]. For quantum error correction, one can use a qWGAN to develop new techniques for constructing quantum error codes or assisting error correction procedures [58–61].

References

[1] Marcello Benedetti, Edward Grant, Leonard Wossnig, and Simone Severini. Adversarial quantum circuit learning for pure state approximation. *New Journal of Physics*, 21(4):043023, 2019.

[2] Pierre-Luc Dallaire-Demers and Nathan Killoran. Quantum generative adversarial networks. *Physical Review A*, 98(1):012324, 2018.

[3] Giacomo Torlai and Roger G Melko. Machine-learning quantum states in the nisq era. *Annual Review of Condensed Matter Physics*, 11:325–344, 2020.

[4] Jun Gao, Lu-Feng Qiao, Zhi-Qiang Jiao, Yue-Chi Ma, Cheng-Qiu Hu, Ruo-Jing Ren, Ai-Lin Yang, Hao Tang, Man-Hong Yung, and Xian-Min Jin. Experimental machine learning of quantum states. *Physical review letters*, 120(24):240501, 2018.

[5] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007.

[6] Andrea Rocchetto, Scott Aaronson, Simone Severini, Gonzalo Carvacho, Davide Poderini, Iris Agresti, Marco Bentivegna, and Fabio Sciarrino. Experimental learning of quantum states. *Science advances*, 5(3):eaau1946, 2019.

[7] Seth Lloyd and Christian Weedbrook. Quantum generative adversarial learning. *Physical review letters*, 121(4):040502, 2018.

[8] Juan Carrasquilla, Giacomo Torlai, Roger G Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, 2019.

[9] Shouvanik Chakrabarti, Huang Yiming, Tongyang Li, Soheil Feizi, and Xiaodi Wu. Quantum wasserstein generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 6781–6792, 2019.

[10] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature communications*, 11(1):1–6, 2020.

[11] Bobak Toussi Kiani, Seth Lloyd, and Reevu Maity. Learning unitaries by gradient descent. *arXiv preprint arXiv:2001.11897*, 2020.

[12] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

[13] Alessandro Bisio, Giulio Chiribella, Giacomo Mauro D'Ariano, Stefano Facchini, and Paolo Perinotti. Optimal quantum learning of a unitary transformation. *Physical Review A*, 81(3):032324, 2010.

[14] Marco Túlio Quintino, Qingxiuxiong Dong, Atsushi Shimbo, Akihito Soeda, and Mio Murao. Reversing unknown quantum transformations: Universal quantum circuit for inverting general unitary operations. *Physical Review Letters*, 123(21):210502, 2019.

[15] Seth Lloyd, Samuel Bosch, Giacomo De Palma, Bobak Kiani, Zi-Wen Liu, Milad Marvian, Patrick Rebentrost, and David M Arvidsson-Shukur. Quantum polar decomposition algorithm. *arXiv preprint arXiv:2006.00841*, 2020.

[16] Jacques Carolan, Masoud Mohseni, Jonathan P Olson, Mihika Prabhu, Changchen Chen, Darius Bunandar, Murphy Yuezhen Niu, Nicholas C Harris, Franco NC Wong, Michael Hochberg, et al. Variational quantum unsampling on a quantum photonic processor. *Nature Physics*, 16(3):322–327, 2020.

[17] Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information*, 5(1):1–9, 2019.

[18] Jin-Guo Liu and Lei Wang. Differentiable learning of quantum circuit born machines. *Physical Review A*, 98(6):062324, 2018.

[19] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The born supremacy: Quantum advantage and training of an ising born machine. *npj Quantum Information*, 6(1):1–11, 2020.

[20] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):1–6, 2018.

[21] Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *arXiv preprint arXiv:2007.14384*, 2020.

[22] Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost-function-dependent barren plateaus in shallow quantum neural networks. *arXiv preprint arXiv:2001.00550*, 2020.

[23] Alexander N Pechen and David J Tannor. Are there traps in quantum control landscapes? *Physical review letters*, 106(12):120402, 2011.

[24] Katharine W Moore and Herschel Rabitz. Exploring constrained quantum control landscapes. *The Journal of chemical physics*, 137(13):134113, 2012.

[25] Giacomo De Palma, Milad Marvian, Dario Trevisan, and Seth Lloyd. The quantum wasserstein distance of order 1. *arXiv preprint arXiv:2009.04469*, 2020.

[26] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[27] Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pages 4666–4677, 2018.

[28] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

[29] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[30] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[32] Ling Hu, Shu-Hao Wu, Weizhou Cai, Yuwei Ma, Xianghao Mu, Yuan Xu, Haiyan Wang, Yipu Song, Dong-Ling Deng, Chang-Ling Zou, et al. Quantum generative adversarial learning in a superconducting quantum circuit. *Science advances*, 5(1):eaav2761, 2019.

[33] Andrea Skolik, Jarrod R McClean, Masoud Mohseni, Patrick van der Smagt, and Martin Leib. Layerwise learning for quantum neural networks. *arXiv preprint arXiv:2006.14904*, 2020.

[34] Fernando GSL Brandao and Krysta M Svore. Quantum speed-ups for solving semidefinite programs. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 415–426. IEEE, 2017.

[35] Joran Van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Quantum sdp-solvers: Better upper and lower bounds. *Quantum*, 4:230, 2020.

[36] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[38] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[41] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[43] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

[44] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. The expressive power of parameterized quantum circuits. *arXiv preprint arXiv:1810.11922*, 2018.

[45] Kunal Sharma, Marco Cerezo, Lukasz Cincio, and Patrick J Coles. Trainability of dissipative perceptron-based quantum neural networks. *arXiv preprint arXiv:2005.12458*, 2020.

[46] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. The quest for a quantum neural network. *Quantum Information Processing*, 13(11):2567–2586, 2014.

[47] Nathan Killoran, Thomas R Bromley, Juan Miguel Arrazola, Maria Schuld, Nicolás Quesada, and Seth Lloyd. Continuous-variable quantum neural networks. *Physical Review Research*, 1(3):033063, 2019.

[48] Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

[49] Patrick Huembeli and Alexandre Dauphin. Characterizing the loss landscape of variational quantum circuits. *arXiv preprint arXiv:2008.02785*, 2020.

[50] Mark Fingerhuth, Tomáš Babej, et al. A quantum alternating operator ansatz with hard and soft constraints for lattice protein folding. *arXiv preprint arXiv:1810.13411*, 2018.

[51] Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, 2019.

[52] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

[53] Xiaozhen Ge, Haijin Ding, Herschel Rabitz, and Re-Bing Wu. Robust quantum control in games: An adversarial learning approach. *Physical Review A*, 101(5):052317, 2020.

[54] Pantita Palittapongarnpim, Peter Wittek, Ehsan Zahedinejad, Shakib Vedaie, and Barry C Sanders. Learning in quantum control: High-dimensional global optimization for noisy quantum dynamics. *Neurocomputing*, 268:116–126, 2017.

[55] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

[56] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology*, 2(4):045001, 2017.

[57] Tyson Jones and Simon C Benjamin. Quantum compilation and circuit optimisation via energy dissipation. *arXiv preprint arXiv:1811.03147*, 2018.

[58] Hendrik Poulsen Nautrup, Nicolas Delfosse, Vedran Dunjko, Hans J Briegel, and Nicolai Friis. Optimizing quantum error correction codes with reinforcement learning. *Quantum*, 3:215, 2019.

[59] Paul Baireuther, Thomas E O'Brien, Brian Tarasinski, and Carlo WJ Beenakker. Machine-learning-assisted correction of correlated qubit errors in a topological code. *Quantum*, 2:48, 2018.

[60] Johannes Bausch and Felix Leditzky. Quantum codes from neural networks. *New Journal of Physics*, 22(2):023005, 2020.

[61] Peter D Johnson, Jonathan Romero, Jonathan Olson, Yudong Cao, and Alán Aspuru-Guzik. Qvector: an algorithm for device-tailored quantum error correction. *arXiv preprint arXiv:1711.02249*, 2017.

[62] Martin Larocca, Esteban A Calzetta, and Diego A Wisniacki. Navigating on quantum control solution subspaces. *arXiv preprint arXiv:2001.05941*, 2020.

[63] Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019.

[64] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D Lukin. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Physical Review X*, 10(2):021067, 2020.

[65] Arthur Pesah, M Cerezo, Samson Wang, Tyler Volkoff, Andrew T Sornborger, and Patrick J Coles. Absence of barren plateaus in quantum convolutional neural networks. *arXiv preprint arXiv:2011.02966*, 2020.

[66] Kishor Bharti and Tobias Haug. Quantum assisted simulator. *arXiv preprint arXiv:2011.06911*, 2020.

[67] James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. *Quantum*, 4:269, 2020.

[68] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[69] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[70] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[71] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in neural information processing systems*, pages 2018–2028, 2017.

[72] Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of wasserstein gans. *arXiv preprint arXiv:1709.08894*, 2017.

[73] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[74] Zefeng Li, Men-Andrin Meier, Egill Hauksson, Zhongwen Zhan, and Jennifer Andrews. Machine learning seismic wave discrimination: Application to earthquake early warning. *Geophysical Research Letters*, 45(10):4773–4779, 2018.

[75] Qi Xuan, Zhuangzhi Chen, Yi Liu, Huimin Huang, Guanjun Bao, and Dan Zhang. Multiview generative adversarial network and its application in pearl classification. *IEEE Transactions on Industrial Electronics*, 66(10):8244–8252, 2018.

[76] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[77] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[78] Zirui Wang, Jun Wang, and Youren Wang. An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. *Neurocomputing*, 310:213–222, 2018.

[79] Abhinav Anand, Jonathan Romero, Matthias Degroote, and Alán Aspuru-Guzik. Experimental demonstration of a quantum generative adversarial network for continuous distributions. *arXiv preprint arXiv:2006.01976*, 2020.

[80] Shahnawaz Ahmed, Carlos Sánchez Muñoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *arXiv preprint arXiv:2008.03240*, 2020.

[81] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020.

[82] Jinfeng Zeng, Yufeng Wu, Jin-Guo Liu, Lei Wang, and Jiangping Hu. Learning and inference on generative adversarial quantum circuits. *Physical Review A*, 99(5):052306, 2019.

[83] Jonathan Romero and Alan Aspuru-Guzik. Variational quantum generators: Generative adversarial quantum machine learning for continuous distributions. *arXiv preprint arXiv:1901.00848*, 2019.

[84] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. Quantum generative adversarial networks for learning and loading random distributions. *npj Quantum Information*, 5(1):1–9, 2019.

[85] Kouhei Nakaji and Naoki Yamamoto. Quantum semi-supervised generative adversarial network for enhanced data classification. *arXiv preprint arXiv:2010.13727*, 2020.

[86] Daniel Herr, Benjamin Obert, and Matthias Rosenkranz. Anomaly detection with variational quantum generative adversarial networks. *arXiv preprint arXiv:2010.10492*, 2020.

[87] Bing Huang, Nadine O Symonds, and O Anatole von Lilienfeld. Quantum machine learning in chemistry and materials. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1883–1909, 2020.

[88] Nikitas Stamatopoulos, Daniel J Egger, Yue Sun, Christa Zoufal, Raban Iten, Ning Shen, and Stefan Woerner. Option pricing using quantum computers. *Quantum*, 4:291, 2020.

[89] Roman Orus, Samuel Mugel, and Enrique Lizaso. Quantum computing for finance: overview and prospects. *Reviews in Physics*, 4:100028, 2019.

[90] Bobak Toussi Kiani, Agnes Villanyi, and Seth Lloyd. Quantum medical imaging algorithms. *arXiv preprint arXiv:2004.02036*, 2020.

[91] Bobak T Kiani, Giacomo De Palma, Dirk Englund, William Kaminsky, Milad Marvian, and Seth Lloyd. Quantum advantage for differential equation analysis. *arXiv preprint arXiv:2010.15776*, 2020.

[92] Xi-Wei Yao, Hengyan Wang, Zeyang Liao, Ming-Cheng Chen, Jian Pan, Jun Li, Kechao Zhang, Xingcheng Lin, Zhehui Wang, Zhihuang Luo, et al. Quantum image processing and its application to edge detection: theory and experiment. *Physical Review X*, 7(3):031041, 2017.

[93] Seth Lloyd. Quantum approximate optimization is computationally universal. *arXiv preprint arXiv:1812.11075*, 2018.

[94] Yuxuan Zhang, Ruizhe Zhang, and Andrew C Potter. Qed driven qaoa for network-flow optimization. *arXiv preprint arXiv:2006.09418*, 2020.

[95] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.

[96] Robert M Parrish, Edward G Hohenstein, Peter L McMahon, and Todd J Martínez. Quantum computation of electronic transitions using a variational quantum eigensolver. *Physical review letters*, 122(23):230401, 2019.

[97] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[98] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[99] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[100] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[101] Anatoly Moiseevich Vershik. Long history of the Monge-Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013.

[102] Donald S Ornstein. An application of ergodic theory to probability theory. *The Annals of Probability*, 1(1):43–58, 1973.

[103] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[104] Eric A Carlen and Jan Maas. An analog of the 2-Wasserstein metric in non-commutative probability under which the Fermionic Fokker–Planck equation is gradient flow for the entropy. *Communications in Mathematical Physics*, 331(3):887–926, 2014.

[105] Eric A Carlen and Jan Maas. Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance. *Journal of Functional Analysis*, 273(5):1810–1869, 2017.

[106] Eric A Carlen and Jan Maas. Non-commutative calculus, optimal transport and functional inequalities in dissipative quantum systems. *Journal of Statistical Physics*, 178(2):319–378, 2020.

[107] Cambyse Rouzé and Nilanjana Datta. Concentration of quantum states from quantum functional and transportation cost inequalities. *Journal of Mathematical Physics*, 60(1):012202, 2019.

[108] Nilanjana Datta and Cambyse Rouzé. Relating relative entropy, optimal transport and Fisher information: A quantum HWI inequality. *Annales Henri Poincaré*, 21:2115–2150, 2020.

[109] Tan Van Vu and Yoshihiko Hasegawa. Geometrical Bounds of the Irreversibility in Markovian Systems. *arXiv preprint arXiv:2005.02871*, 2020.

[110] Giacomo De Palma and Stefan Huber. The conditional Entropy Power Inequality for quantum additive noise channels. *Journal of Mathematical Physics*, 59(12):122201, 2018.

[111] Li Gao, Marius Junge, and Nicholas LaRacuente. Fisher information and logarithmic sobolev inequality for matrix-valued functions. *Annales Henri Poincaré*, 21(11):3409–3478, 2020.

[112] Yongxin Chen, Tryphon T Georgiou, Lipeng Ning, and Allen Tannenbaum. Matricial Wasserstein-1 distance. *IEEE control systems letters*, 1(1):14–19, 2017.

[113] Ernest K Ryu, Yongxin Chen, Wuchen Li, and Stanley Osher. Vector and matrix optimal mass transport: theory, algorithm, and applications. *SIAM Journal on Scientific Computing*, 40(5):A3675–A3698, 2018.

[114] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: a quantum mechanical approach. *IEEE Transactions on Automatic Control*, 63(8):2612–2619, 2018.

[115] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Wasserstein geometry of quantum states and optimal transport of matrix-valued measures. In *Emerging Applications of Control and Systems Theory*, pages 139–150. Springer, 2018.

[116] Julián Agredo. A Wasserstein-type distance to measure deviation from equilibrium of quantum Markov semigroups. *Open Systems & Information Dynamics*, 20(02):1350009, 2013.

[117] J Agredo. On exponential convergence of generic quantum Markov semigroups in a Wasserstein-type distance. *International Journal of Pure and Applied Mathematics*, 107(4):909–925, 2016.

[118] Kazuki Ikeda. Foundation of quantum optimal transport and applications. *Quantum Information Processing*, 19(1):25, 2020.

[119] François Golse, Clément Mouhot, and Thierry Paul. On the mean field and classical limits of quantum mechanics. *Communications in Mathematical Physics*, 343(1):165–205, 2016.

[120] Emanuele Caglioti, François Golse, and Thierry Paul. Towards optimal transport for quantum densities. preprint, Dec 2018.

[121] François Golse. The quantum N-body problem in the mean-field and semiclassical regime. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170229, 2018.

[122] François Golse and Thierry Paul. The Schrödinger equation in the mean-field and semiclassical regime. *Archive for Rational Mechanics and Analysis*, 223(1):57–94, 2017.

[123] François Golse and Thierry Paul. Wave packets and the quadratic Monge–Kantorovich distance in quantum mechanics. *Comptes Rendus Mathematique*, 356(2):177–197, 2018.

[124] Emanuele Caglioti, François Golse, and Thierry Paul. Quantum Optimal Transport is Cheaper. *Journal of Statistical Physics*, 2020.

[125] Giacomo De Palma and Dario Trevisan. Quantum optimal transport with quantum channels. *arXiv preprint arXiv:1911.00803*, 2019.

[126] Rocco Duvenhage and Machiel Snyman. Balance between quantum Markov semigroups. *Annales Henri Poincaré*, 19(6):1747–1786, 2018.

[127] J Agredo and Franco Fagnola. On quantum versions of the classical Wasserstein distance. *Stochastics*, 89(6-7):910–922, 2017.

[128] Karol Zyczkowski and Wojeciech Slomczynski. The Monge distance between quantum states. *Journal of Physics A: Mathematical and General*, 31(45):9095, 1998.

[129] Karol Zyczkowski and Wojciech Slomczynski. The Monge metric on the sphere and geometry of quantum states. *Journal of Physics A: Mathematical and General*, 34(34):6689, 2001.

[130] Ingemar Bengtsson and Karol Życzkowski. *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge University Press, 2017.

[131] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[132] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

[133] Michael Mathieu and Yann LeCun. Fast approximation of rotations and hessians matrices. *arXiv preprint arXiv:1404.7195*, 2014.

[134] Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, and Marin Soljačić. Tunable efficient unitary neural networks (eunn) and their application to rnns. In *International Conference on Machine Learning*, pages 1733–1741. PMLR, 2017.

[135] Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. *Proceedings of machine learning research*, 97:1517, 2019.

[136] William R Clements, Peter C Humphreys, Benjamin J Metcalf, W Steven Kolthammer, and Ian A Walmsley. Optimal design for universal multiport interferometers. *Optica*, 3(12):1460–1465, 2016.

[137] Yichen Shen, Nicholas C Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, et al. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441, 2017.

[138] Guillaume Verdon, Michael Broughton, and Jacob Biamonte. A quantum algorithm to train neural networks using low-depth circuits. *arXiv preprint arXiv:1712.05304*, 2017.

[139] Zhihui Wang, Stuart Hadfield, Zhang Jiang, and Eleanor G Rieffel. Quantum approximate optimization algorithm for maxcut: A fermionic view. *Physical Review A*, 97(2):022304, 2018.

[140] Mark Hodson, Brendan Ruck, Hugh Ong, David Garvin, and Stefan Dulman. Portfolio rebalancing experiments using the quantum alternating operator ansatz. *arXiv preprint arXiv:1911.05296*, 2019.

[141] Nicholas Chancellor. Domain wall encoding of discrete variables for quantum annealing and qaoa. *Quantum Science and Technology*, 4(4):045004, 2019.

[142] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, M Sohaib Alam, Shahnawaz Ahmed, Juan Miguel Arrazola, Carsten Blank, Alain Delgado, Soran Jahangiri, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

[143] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat,

Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[144] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

Appendix A Related Works

A.1 Loss landscape in quantum machine learning

Prior research has theoretically and numerically analyzed the typical properties of loss landscapes in quantum machine learning and control settings. Notably, for commonly used cost functions, prior work has proved and numerically shown the existence of "barren plateaus" characterized by exponentially decaying gradients for large depth quantum parameterized circuits [20], cost functions with global observables [22], and noisy circuits [21]. Furthermore, research in quantum control theory has identified the presence of traps when the control landscape is constrained [23, 24, 62]. Various attempts have been made to potentially avoid these roadblocks. These include methods for initializing circuit parameters [63, 64], algorithms for layer-wise training [33], and choosing ansatzes that can avoid decaying gradients [65, 66]. More pertinent to our work, various authors have chosen loss metrics or designed gradient-based algorithms that help avoid issues with barren plateaus. This includes Ref. [22], which considers local operator loss functions, Ref. [49], which includes second order derivatives in optimizing the loss function to better navigate flat loss landscapes, and Ref. [67], which constructed an algorithm to optimize over the Fubini-study metric tensor.

A.2 Classical and quantum generative adversarial networks

Generative models, as their name indicates, aim to generate a target object or produce samples from a target distribution by approximating the given target through a learning procedure. In the quantum setting, one popular variant for generative models are Born machines whose cost function is measured by comparing a target classical distribution to the sample distribution of a measurement from a variational quantum circuit [17–19]. Ref. [17] considers the classical EM distance in their evaluation of the cost function which compares the sampled distribution of the quantum computer to the target distribution.

One commonly used generative algorithm is the generative adversarial network (GAN), a classical algorithm first introduced in [31]. Most relevant to the current work, Ref. [26] constructed the first classical Wasserstein GAN employing an earth mover's distance. Later work improved upon the stability and training of the original Wasserstein GAN by, for example, constructing improved discriminators and generators [30, 68, 69], progressively adding layers during training [70], and employing various regularization techniques [71–73]. In the classical literature, GANs have been extensively used in many real-world applications [74–78].

In the quantum setting, quantum GANs were first proposed by Refs. [2, 7]. Simple experiments were performed showing the power of quantum GANs in learning quantum data for relatively small systems that can be simulated or experimentally analyzed [1, 32, 79–81]. Hybrid classical-quantum GANs, fully classical in the discriminator, generator, and/or loss function, were proposed in Refs. [82–85]. Ref. [9] proposed a version of a quantum Wasserstein GAN (qWGAN), though the employed earth mover's distance is unitarily invariant (see Appendix D for details). Ref. [86] also proposed a qWGAN structure with a classical discriminator and the classical EM distance as their loss function. Our work differs from

both of these prior qWGAN papers in that it implements the first qWGAN with a quantum EM distance. Some early experimental demonstrations of quantum GANs have also been performed on various different systems [32, 79, 87, 88].

A.3 Applications of quantum machine learning

Most of the work in quantum machine learning has focused on finding useful applications for quantum machine learning. These include applications in finance [88, 89], chemistry [87], and post-processing quantum outputs [90–92]. Beyond quantum GANs, there has been a focus in recent years on developing near term quantum algorithms potentially implementable on quantum computers with around 100 qubits. Among the most promising candidates include the quantum approximate optimization algorithm [52,93,94], the variational quantum eigensolver [95, 96], and quantum GANs discussed earlier.

Appendix B The Classical Earth Mover's Distance

The classical earth mover's (EM) distance, also called Monge-Kantorovich distance, is a distance between probability distributions on a metric space which dates back to Monge [97] and has its roots in the theory of optimal mass transport. Let p, q be probability distributions on the metric space $\mathcal{X}$, which for simplicity we will assume to be finite, and let d be the distance on $\mathcal{X}$. Following the Kantorovich's formulation of the EM distance [98], we define the set of the *couplings* between p and q as the set of the probability distributions on two copies of $\mathcal{X}$ with marginals equal to p and q, respectively. In the interpretation of mass transport, p and q are considered as distributions of a unit amount of mass, and any coupling π prescribes a plan to transform the distribution p into the distribution q, in the sense that $\pi(x, y)$ is the amount of mass that is moved from x to y. Assuming that the cost of moving a unit of mass from x to y is equal to $d(x, y)$, the cost of the coupling π is equal to $\sum_{x,\, y\in\mathcal{X}} \pi(x, y)\, d(x, y)$, *i.e.*, to the expectation value of the distance with respect to π. The EM distance between p and q is given by the minimum cost among all the couplings between p and q. The EM distance has been generalized to a transport cost equal to a power of d, leading to the family of the Wasserstein distances of order α, of which the $\alpha = 1$ case recovers the EM distance. The exploration of the Wasserstein distances has led to the creation of an extremely fruitful field in mathematical analysis, with applications ranging from differential geometry and partial differential equations to machine learning [29, 99–101].

The EM distance can be considered as a generalization of the total variation distance. Indeed, the EM distance recovers the total variation distance when the distance d on $\mathcal{X}$ is the trivial distance for which all the elements of $\mathcal{X}$ are equivalent, *i.e.*, $d(x, y) = 1$ for any $x \neq y \in \mathcal{X}$.

When $\mathcal{X}$ is a set of the strings of n bits, the natural choice for d is the Hamming distance, given by the number of different bits. In this case, the EM distance is also known as Ornstein's $\bar{d}$ distance [102].

Appendix C Quantum Mechanics and Qubits

Any quantum system has an associated Hilbert space. If the Hilbert space has finite dimension N, it is always isomorphic to $\mathbb{C}^N$. For the sake of simplicity, we restrict our discussion to this case.

We denote a column vector in $\mathbb{C}^N$ with $|\cdot\rangle$, where $\cdot$ is a label for the vector. We will mostly consider vectors with unit norm. For any $|\psi\rangle \in \mathbb{C}^N$, we denote with $\langle\psi| \in \left(\mathbb{C}^N\right)^*$ the row vector whose entries are the complex conjugates of the entries of $|\psi\rangle$. Following the usual rule for matrix multiplication, $\langle\cdot|\cdot\rangle$ denotes the canonical Hermitian inner product of $\mathbb{C}^N$, defined to be antilinear in the first entry and linear in the second.

A *quantum state* is the quantum counterpart of a probability distribution on a set of N elements, and is a positive semidefinite Hermitian matrix in $\mathbb{C}^{N\times N}$ with unit trace. A quantum state is *pure* if it cannot be expressed as a nontrivial convex combination of quantum states. This is the case iff the quantum state is an orthogonal projector with rank one, *i.e.*, if it can be expressed as $|\psi\rangle\langle\psi|$ for some unit vector $|\psi\rangle \in \mathbb{C}^N$. With some abuse of notation, we call also the unit vectors in $\mathbb{C}^N$ quantum states, formally meaning the associated orthogonal projectors. Similarly, we call the inner product between two pure quantum states the inner product between the associated unit vectors. A quantum state is called *mixed* if it is not pure. Two quantum states are called *orthogonal* if the corresponding supports are orthogonal. Any mixed quantum state can be expressed as a convex combination of mutually orthogonal pure quantum states.

An *observable* is the quantum counterpart of a real-valued function on a set of N elements, and is given by an $N \times N$ Hermitian matrix. The expectation value of the observable H on the quantum state ρ is given by $\mathrm{Tr}\left[\rho\, H\right]$.

The Hilbert space associated to a composite quantum system is the tensor product of the Hilbert spaces associated to each subsystem. Let ρ be a quantum state of the composite quantum system with Hilbert space $\mathbb{C}^{N_1} \otimes \mathbb{C}^{N_2}$, *i.e.*, a Hermitian matrix in $\mathbb{C}^{N_1\times N_1} \otimes \mathbb{C}^{N_2\times N_2}$. We denote with ρ_1 the *marginal* state of ρ on the first subsystem, *i.e.*, the quantum state in $\mathbb{C}^{N_1\times N_1}$ such that $\mathrm{Tr}\left[\rho_1\, H\right] = \mathrm{Tr}\left[\rho\,(H \otimes \mathbb{I}_{N_2})\right]$ for any quantum observable H of $\mathbb{C}^{N_1}$. ρ_1 is equal to the partial trace of ρ over the second subsystem: $\rho_1 = \mathrm{Tr}_2\rho$.

In this paper, we focus on a quantum system composed of n qubits. A *qubit* is the quantum system associated to the Hilbert space $\mathbb{C}^2$. We denote with $|0\rangle, |1\rangle$ the vectors of its canonical basis, which is also called the computational basis. The Hilbert space of n qubits is $\left(\mathbb{C}^2\right)^{\otimes n}$, and is isomorphic to $\mathbb{C}^N$ with $N = 2^n$. The computational basis of $\left(\mathbb{C}^2\right)^{\otimes n}$ is $\{|x_1\rangle \otimes \ldots \otimes |x_n\rangle : x \in \{0,1\}^n\}$. By the sake of a simpler notation, we denote each vector $|x_1\rangle \otimes \ldots \otimes |x_n\rangle$ with $|x\rangle$, and we set $|0\rangle^{\otimes n} = |0_n\rangle$, $|1\rangle^{\otimes n} = |1_n\rangle$. We denote with $\mathcal{O}_n$ the set of the observables of $\left(\mathbb{C}^2\right)^{\otimes n}$. We say that a linear operator on $\left(\mathbb{C}^2\right)^{\otimes n}$ acts on the i-th qubit if it is equal to a 2×2 matrix acting on the i-th qubit tensored with the identity operator acting on the remaining $n-1$ qubits. The definition of a linear operator acting on a subset of qubits is analogous.

Perhaps the most important observables used and studied in quantum computation are the Pauli matrices. Together with the identity matrix, the Pauli matrices shown below form

a basis for the observables on $\mathbb{C}^2$ (*i.e.,* one qubit).

$$\sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{18}$$

$$\sigma_Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{19}$$

$$\sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{20}$$

A single Pauli observable can act as a measurement on one qubit; however, multiple qubits can be measured by a "Pauli string" represented by a set of Pauli matrices placed in tensor product form (*e.g.,* $\sigma_X \otimes \sigma_Y \otimes I \otimes \sigma_X$ or equivalently the string 'XYIX').

Pauli operators are often used as parameterized quantum gates. In their parameterized form:

$$R_P(t) = e^{-it\sigma_P/2} = \cos\left(\frac{t}{2}\right) I - i \sin\left(\frac{t}{2}\right) \sigma_P \tag{21}$$

where subscript $P \in \{X, Y, Z\}$ indicates the specific Pauli operator chosen.

In our exposition, we outline the computations performed on a quantum computer as quantum circuits, which are models representing a computation as a sequence of reversible quantum gates and measurement operators. Quantum circuits contain n-bit registers and the sequence of gates are applied accordingly to the qubits in the register. For further details on how to read quantum circuits, the reader is referred to the book [103].

Appendix D Quantum Generalizations of Wasserstein Distances

Several quantum generalizations of optimal transport distances have been proposed. One line of research by Carlen, Maas, Datta and Rouzé [104–109] defines a quantum Wasserstein distance of order 2 from a Riemannian metric on the space of quantum states based on a quantum analog of a differential structure. This quantum Wasserstein distance is intimately linked to both entropy and Fisher information [108], and has led to determine the rate of convergence of the quantum Ornstein-Uhlenbeck semigroup [105, 110]. Exploiting their quantum differential structure, Refs. [106, 107, 111] also define a quantum generalization of the Lipschitz constant and of the earth mover's distance. Alternative definitions of quantum earth mover's distances based on a quantum differential structure are proposed in Refs. [112–115]. Refs. [116–118] propose quantum earth mover's distances based on a distance between the vectors of the canonical basis.

Another line of research by Golse, Mouhot, Paul and Caglioti [119–124] arose in the context of the study of the semiclassical limit of quantum mechanics and defines a family of quantum Wasserstein distances of order 2 built on the notion of couplings. A coupling between the quantum states ρ and σ of $\mathbb{C}^N$ is a quantum state Π of $\left(\mathbb{C}^N\right)^{\otimes 2}$ whose marginal

states on the first and on the second subsystems are equal to ρ and σ, respectively. The transport cost of the coupling Π is $\mathrm{Tr}\left[\Pi\, C\right]$, where C is a suitable positive semidefinite $N^2 \times N^2$ cost matrix. Different choices of C will lead to different distances. The square distance between ρ and σ is defined as the minimum cost among all the couplings between ρ and σ. Refs. [119–124] consider the case of a quantum harmonic oscillator, which is actually infinite dimensional, and choose as cost matrix the quantum analog of the square Euclidean distance:

$$C = \left(Q_1 - Q_2\right)^2 + \left(P_1 - P_2\right)^2 \,, \tag{22}$$

where $Q_{1,2}$ and $P_{1,2}$ are the position and momentum operators of the two subsystems, respectively. However, the resulting distance has the undesirable property that the distance between a quantum state and itself may not be zero. Ref. [9] notices that the distance between any quantum state and itself is zero whenever the support of the cost matrix C is contained in the antisymmetric subspace with respect to the swap of the two subsystems of $\left(\mathbb{C}^N\right)^{\otimes 2}$. Therefore, Ref. [9] chooses the orthogonal projector onto the antisymmetric subspace as cost matrix, and employs the resulting distance as a cost function for quantum GANs. We stress that this distance is unitarily invariant. Indeed, for any $N \times N$ unitary matrix U, if Π is a coupling between the quantum states ρ and σ, then $U^{\otimes 2}\, \Pi\, U^{\dagger \otimes 2}$ is a coupling between $U\, \rho\, U^\dagger$ and $U\, \sigma\, U^\dagger$, and these two couplings have the same cost since the projector onto the antisymmetric subspace commutes with $U^{\otimes 2}$. Moreover, the only coupling between the pure quantum states $|\psi\rangle$ and $|\phi\rangle$ is the product state $\Pi = |\psi\rangle\langle\psi| \otimes |\phi\rangle\langle\phi|$, whose cost is equal to $\left(1 - |\,\langle\phi|\psi\rangle\,|^2\right)/2$. Therefore, the distance between pure quantum states is a function of their overlap.

Ref. [125] proposes another quantum Wasserstein distance of order 2 based on couplings, with the property that each quantum coupling is associated to a quantum channel. The relation between quantum couplings and quantum channels in the framework of von Neumann algebras has been explored in [126]. The problem of defining a quantum earth mover's distance through quantum couplings has been explored in Ref. [127].

The quantum Wasserstein distance between two quantum states can be defined as the classical Wasserstein distance between the probability distributions of the outcomes of an informationally complete measurement performed on the states, which is a measurement whose probability distribution completely determines the state. This definition has been explored for Gaussian quantum systems with the heterodyne measurement in Refs. [128–130].

Appendix E Quantum EM distance and Lipschitz constant

The quantum EM distance between the quantum states of n qubits ρ and σ is given by the following maximization semidefinite program [25, Definition 6]:

$$\|\rho - \sigma\|_{EM} = \frac{1}{2} \max \left(\sum_{i=1}^{n} \|X_i\|_1 : X_i \in \mathcal{O}_n,\ \mathrm{Tr}_i X_i = 0\ \forall\, i = 1, \ldots, n,\ \sum_{i=1}^{n} X_i = \rho - \sigma \right). \tag{23}$$

Appendix F Toy Model Details

Our toy model (subsection 2.2) analyzes the learnability of the GHZ state when using a loss function either corresponding to fidelity (function of inner product between target GHZ state and generated state) or the quantum EM distance. For optimizing over the fidelity, we have a loss function, copied below, that is easily evaluated as a function of $\boldsymbol{\theta}$.

$$F = |\langle GHZ_n | \psi(\boldsymbol{\theta}) \rangle|^2 = \left(\frac{\cos(\theta_1)}{\sqrt{2}} + \frac{\prod_{i=1}^{n} \sin(\theta_i)}{\sqrt{2}} \right)^2 \tag{24}$$

Here, to perform optimization, we simply perform gradient based updates on the parameters θ_i in the equation above. For all experiments, the Adam optimizer was used to perform gradient updates with a learning rate of 0.2 [131]. For each simulation, 100000 steps of optimization were performed before stopping. Learning is considered successful if $F < 0.02$.

In our toy model, to efficiently approximate the quantum EM distance, we construct a loss function $\tilde{D}_{EM} \approx D_{EM}\left(|\psi(\boldsymbol{\theta})\rangle \langle\psi(\boldsymbol{\theta})|, |GHZ_n\rangle \langle GHZ_n|\right)$ which takes the maximum over $O(n)$ expectations of Pauli operators. We first note that the state $|\psi(\boldsymbol{\theta})\rangle$ is spanned by up to $n+1$ computational basis states.

$$\begin{aligned} |\psi(\boldsymbol{\theta})\rangle &= \cos\theta_1 |0_n\rangle + i \sin\theta_1 \cos\theta_2 |1\rangle |0_{n-1}\rangle - \sin\theta_1 \sin\theta_2 \cos\theta_3 |1_2\rangle |0_{n-2}\rangle + \ldots \\ &= \cos\theta_1 |0_n\rangle + \sum_{k=1}^{n-1} i^k \left[\prod_{j=1}^{k} \sin\theta_j \right] \cos\theta_{k+1} |1_k\rangle |0_{n-k}\rangle + i^n \left[\prod_{j=1}^{n} \sin\theta_j \right] |1_n\rangle \end{aligned} \tag{25}$$

We can write the state above in a vector of length $n+1$ only including the terms in the above span:

$$|\psi(\boldsymbol{\theta})\rangle = \begin{bmatrix} \cos\theta_1 \\ i \sin\theta_1 \cos\theta_2 \\ \vdots \\ i^n \prod_{j=1}^{n} \sin\theta_j \end{bmatrix}, \tag{26}$$

where the above vector can be easily stored in the memory of a classical computer.

To calculate $\tilde{D}_{EM}$, we first measure the expectation of $|\psi(\boldsymbol{\theta})\rangle$ with respect to the following $2n$ Pauli operators P_i:

$$\begin{aligned} P_i \in \{&\sigma_Z^{(1)},\ \sigma_Z^{(2)}, \ldots,\ \sigma_Z^{(n)}, \\ &\sigma_Y^{(1)},\ \sigma_X^{(1)} \otimes \sigma_X^{(2)},\ \sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \sigma_Y^{(3)}, \ldots,\ \sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \cdots \otimes \sigma_X^{(n)}\}, \end{aligned} \tag{27}$$

which is equivalent to the complete set of single qubit Pauli Z operators combined with a multi-qubit Pauli operator for each qubit k consisting of the Pauli X operator or Pauli Y operator acting on qubit k if k is even or odd respectively (to handle relative phases) and Pauli X operators acting on all qubits $j < k$. In the above, we use the notation $\sigma_L^{(i)}$ to indicate Pauli $L \in \{X, Y, Z\}$ acting on qubit i.

Since as mentioned earlier, $|\psi(\boldsymbol{\theta})\rangle$ is written compactly in vector form, expectations for each of the above operators can be efficiently evaluated using a classical computer. As discussed in section 3, an optimal Hamiltonian whose expectation approximates the quantum EM distance can be efficiently constructed as a parameterized sum of the above operators. Since all Pauli Z operators act on individual qubits, $\tilde{D}_{EM}$ can be calculated as the maximum amongst the following $n+1$ parameterized sums of expectations of operators:

$$\begin{aligned} \tilde{D}_{EM} = \max\Big\{ &\big|\mathbb{E}[\sigma_Z^{(1)}]\big| + \big|\mathbb{E}[\sigma_Z^{(2)}]\big| + \cdots + \big|\mathbb{E}[\sigma_Z^{(n)}]\big|, \\ &\big|\mathbb{E}[\sigma_Y^{(1)}]\big| + \big|\mathbb{E}[\sigma_Z^{(2)}]\big| + \cdots + \big|\mathbb{E}[\sigma_Z^{(n)}]\big|, \\ &\big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_Y^{(2)}]\big| + \big|\mathbb{E}[\sigma_Z^{(3)}]\big| + \cdots + \big|\mathbb{E}[\sigma_Z^{(n)}]\big|, \\ &\ldots, \\ &\big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \cdots \otimes \sigma_X^{n}]\big| \Big\}, \end{aligned} \tag{28}$$

where $\mathbb{E}[\cdot]$ indicates the difference in expectation of the operator $\cdot$ on the generated state versus the target GHZ state. For faster simulation, we actually consider the maximum over a simpler set of operators that is equally effective at learning the GHZ state:

$$\begin{aligned} \tilde{D}_{EM} = \max\Big\{ &\big|\mathbb{E}[\sigma_Z^{(1)}]\big| + \big|\mathbb{E}[\sigma_Z^{(2)}]\big| + \cdots + \big|\mathbb{E}[\sigma_Z^{(n)}]\big|, \\ &\big|\mathbb{E}[\sigma_Y^{(1)}]\big|, \\ &\big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_Y^{(2)}]\big|, \\ &\ldots, \\ &\big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \cdots \otimes \sigma_X^{n}]\big| \Big\}. \end{aligned} \tag{29}$$

Using the equation for $\tilde{D}_{EM}$ above, gradient updates can efficiently be performed on the parameters of the circuit. As with the fidelity loss function, we perform optimization with the Adam optimizer at a learning rate of 0.2 [131]. Only up to 10000 steps of optimization

were performed since convergence was almost always achieved within about 1000 steps. Learning is considered successful if $|\langle GHZ_n|\psi(\boldsymbol{\theta})\rangle|^2 > 0.98$ after the optimization. Success was achieved in virtually all instances when using $\tilde{D}_{EM}$.

Appendix G Proof of Proposition 1

For any $k = 0, \ldots, n-1$, let

$$\Delta_k = |\Psi_k\rangle\langle\Psi_k| - |GHZ_n\rangle\langle GHZ_n|\,, \tag{30}$$

and let $\mathcal{D}_1$ be the completely dephasing channel acting on the first qubit. From [25, Proposition 2], the quantum EM distance is contractive with respect to a quantum channel acting on a single qubit. We then have on the one hand

$$\|\Delta_k\|_{EM} \geq \|\mathcal{D}_1(\Delta_k)\|_{EM} = \left\| |1_k\rangle\langle 1_k| \otimes \frac{|0_{n-k}\rangle\langle 0_{n-k}| - |1_{n-k}\rangle\langle 1_{n-k}|}{2} \right\|_{EM} = \frac{n-k}{2}\,. \tag{31}$$

On the other hand, we have

$$\begin{aligned} \|\Delta_k\|_{EM} &\leq \|\mathcal{D}_1(\Delta_k)\|_{EM} + \|\Delta_k - \mathcal{D}_1(\Delta_k)\|_{EM} = \frac{n-k}{2} + \frac{1}{2}\|\Delta_k - \mathcal{D}_1(\Delta_k)\|_1 \\ &= \frac{n-k}{2} + \frac{1}{2}\begin{cases} 1 & k = 0 \\ \sqrt{2} & k = 1, \ldots, n-1 \end{cases}, \end{aligned} \tag{32}$$

where the first equality follows from [25, Proposition 6], stating that $\|X\|_{EM} = \|X\|_1/2$ for any $X \in \mathcal{O}_n$ with $\mathrm{Tr}_1 X = 0$. The claim follows.

Appendix H Bias towards local operators

Here we prove that the optimal Hamiltonian for the maximization problem (15) contains only terms with few qubits whenever all the coefficient c_j associated to single Pauli operators are $\Omega(1)$.

Proposition 2. *Let $w^* : \{I, X, Y, Z\}^n \to \mathbb{R}$ be the set of parameters that achieve the maximum in (15), and let*

$$a = \min_{i=1,\ldots,n} \max_{P=X,Y,Z} \left| \mathrm{Tr}\left[(G - \rho_{\mathrm{tar}})\, \sigma_P^{(i)} \right] \right|\,. \tag{33}$$

Then, $w^*_{P_1 \ldots P_n} = 0$ for any $P_1, \ldots, P_n \in \{I, X, Y, Z\}$ such that

$$|c_{P_1 \ldots P_n}| < a\, |\{i = 1, \ldots, n : P_i \neq I\}|\,. \tag{34}$$

In particular, $w^*_{P_1 \ldots P_n} = 0$ for any Pauli string that acts nontrivially on more than $2/a$ qubits.

Proof. The maximization problem (15) is a linear program with dual

$$\min_{z\in\mathbb{R}^n_{\geq 0}} \sum_{i=1}^{n} z_i \quad : \quad |c_{P_1\ldots P_n}| \leq \sum_{i\in[n]:P_i\neq I} z_i \quad \forall\, P_1,\ldots,P_n \in \{I,X,Y,Z\}\,. \tag{35}$$

Let $z^* \in \mathbb{R}^n_{\geq 0}$ achieve the minimum in (35). For any $P_1,\ldots,P_n \in \{I,X,Y,Z\}^n$ such that $w^*_{P_1\ldots P_n} \neq 0$ we have

$$|c_{P_1\ldots P_n}| = \sum_{i\in[n]:P_i\neq I} z^*_i\,. \tag{36}$$

From (35) we have $a \leq z^*_i$ for any $i = 1,\ldots,n$. Let $P_1,\ldots,P_n \in \{I,X,Y,Z\}$ satisfy (34), and let us assume that $w^*_{P_1\ldots P_n} \neq 0$. We get from (36)

$$|c_{P_1\ldots P_n}| = \sum_{i\in[n]:P_i\neq I} z^*_i \geq a\,|\{i\in[n] : P_i \neq I\}|\,, \tag{37}$$

which contradicts (34), and the claim follows. □

Appendix I Gradients of qWGAN

For the generic version of our generator (equation (17)), optimization is performed over probability parameters p_i and gate parameters in each unitary U_i. The generator optimizes the parameters θ to minimize $\mathrm{Tr}\,[G(\theta)\,H]$, where H is the Hamiltonian provided by the discriminator. The following Proposition 3 proves that the gradient of $\mathrm{Tr}\,[G(\theta)\,H]$ coincides with the gradient of the EM distance between $G(\theta)$ and ρ_{tar} if H is the optimal Hamiltonian that achieves the EM distance in (4). Therefore, our learning algorithm decreases the EM distance between $G(\theta)$ and ρ_{tar}. The proof is in subsection I.1.

Proposition 3. *For any target quantum state σ and any parametric family of quantum states $\rho(t)$, $0 \leq t \leq T$ that is differentiable in $t = 0$,*

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\,\|\rho(t)-\sigma\|_{EM}\Big|_{t=0} = \\ &\max\left(\mathrm{Tr}\,[\rho'(0)\,H] : H\in\mathcal{O}_n\,,\ \|H\|_L \leq 1\,,\ \mathrm{Tr}\,[(\rho(0)-\sigma)\,H] = \|\rho(0)-\sigma\|_{EM}\right)\,.\end{aligned} \tag{38}$$

If $\rho(t)$ admits a differentiable extension to negative values of t, (38) provides the right derivative of $\|\rho(t)-\sigma\|_{EM}$, which can be different from the left derivative if the max *in (38) is nontrivial.*

For parameters p_i, the gradient of $\mathrm{Tr}\,[G(\theta)\,H]$ can be evaluated using U_i:

$$\frac{\partial D_{EM}}{\partial p_i} = \mathrm{Tr}(U_i\rho_0 U_i^\dagger H_{\max})\,, \tag{39}$$

where $H_{\max}$ is the optimal Hamiltonian outputted by the discriminator (equation (16)). For gate parameters, we can use standard techniques [132] for evaluating gradients with respect to gate parameters.

I.1 Proof of Proposition 3

On the one hand, we have for any H as in (38)

$$\liminf_{t\to 0^+} \frac{\|\rho(t)-\sigma\|_{EM} - \|\rho(0)-\sigma\|_{EM}}{t} \geq \liminf_{t\to 0^+} \operatorname{Tr}\left[\frac{\rho(t)-\rho(0)}{t} H\right] = \operatorname{Tr}\left[\rho'(0)\, H\right] . \quad (40)$$

On the other hand, for any $0 < t < T$, let $H(t) \in \mathcal{O}_n$ be traceless and such that $\|H(t)\|_L \leq 1$ and $\operatorname{Tr}\left[(\rho(t)-\sigma)\, H(t)\right] = \|\rho(t)-\sigma\|_{EM}$. We have

$$\limsup_{t\to 0^+} \frac{\|\rho(t)-\sigma\|_{EM} - \|\rho(0)-\sigma\|_{EM}}{t} \leq \limsup_{t\to 0^+} \operatorname{Tr}\left[\frac{\rho(t)-\rho(0)}{t} H(t)\right] . \quad (41)$$

Let $t_k \downarrow 0$ be a sequence that achieves the lim sup in the right-hand side of (41) and such that

$$\lim_{k\to\infty} H(t_k) = H_0 \in \mathcal{O}_n . \quad (42)$$

We have

$$\begin{aligned} \|H_0\|_L &= \lim_{k\to\infty} \|H(t_k)\|_L \leq 1 , \\ \operatorname{Tr}\left[(\rho(0)-\sigma)\, H_0\right] &= \lim_{k\to\infty} \operatorname{Tr}\left[(\rho(t_k)-\sigma)\, H(t_k)\right] = \lim_{k\to\infty} \|\rho(t_k)-\sigma\|_{EM} = \|\rho(0)-\sigma\|_{EM} , \end{aligned} \quad (43)$$

and

$$\limsup_{t\to 0^+} \frac{\|\rho(t)-\sigma\|_{EM} - \|\rho(0)-\sigma\|_{EM}}{t} \leq \lim_{k\to\infty} \operatorname{Tr}\left[\frac{\rho(t_k)-\rho(0)}{t_k} H(t_k)\right] = \operatorname{Tr}\left[\rho'(0)\, H_0\right] , \quad (44)$$

and the claim follows.

Appendix J Additional Simulations and Figures

J.1 Learning the GHZ state

The analysis in subsection 5.1 showed that the qWGAN is especially effective at learning the GHZ state. In addition to the results shown in subsection 5.1, Figure 5 shows the typical dynamics of learning the GHZ state of 4, 8, and 12 qubits. In all cases, the GHZ state is learned within 1000 steps of optimization.

J.2 Teacher-student learning

Supplementary to the results in subsection 5.2, we include Figure 6 which shows the typical profile of learning in the teacher-student setup. In almost all instances, learning of the state generated by the teacher circuit was achieved.

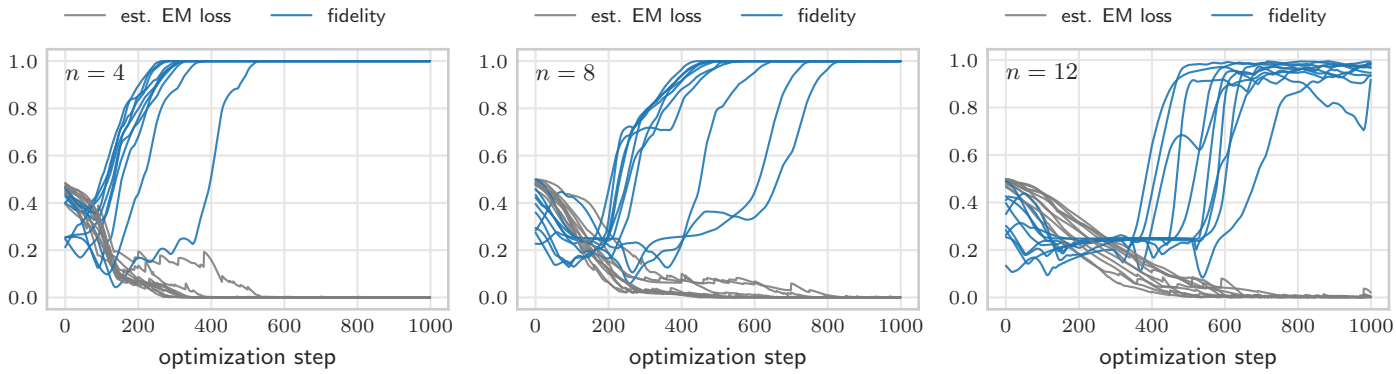


Figure 5: The qWGAN consistently generates the GHZ state in simulations with circuits of 4, 8, and 12 qubits. Estimated EM loss (quantum EM distance estimated by active operators) is also plotted in above chart, normalized by dividing by the number of qubits. Plots contain one line for each of 10 simulations for each circuit size.

J.3 Gradients of qWGAN vs. conventional GANs

Supplementary to Figure 3b, we include further details of the gradients of the quantum EM loss metric and its comparison to a inner product loss metric in Figure 7. As a reminder, the inner product loss metric is $F = 1 - |\langle \phi_{\text{tar}} | \phi(\boldsymbol{\theta}) \rangle|^2$.

J.4 Butterfly circuit learning

In this section, we consider learning the parameters of a "butterfly" circuit which constructs interactions between all qubits in $O(\log_2 n)$ layers. The general form of this circuit is shown in Appendix K and is motivated by prior work in classical machine learning and photonics where similar parameterizations of unitary transformations produced interesting results [133–137]. Here, the generator takes the form of r_{gen} copies of the parameterized butterfly circuit. The generator aims to learn a target density matrix ρ_{tar} of rank r_{tar} which is generated from a circuit of the same form as the generator but with randomly chosen parameters. In other words,

$$\rho_{\text{tar}} = \frac{1}{r_{\text{tar}}} \sum_{i=1}^{r_{\text{tar}}} U_b(\theta_{\text{ran}}^{(i)}) \rho_0 U_b(\theta_{\text{ran}}^{(i)})^\dagger \tag{45}$$

where $U_b(\theta_{\text{ran}}^{(i)})$ is the unitary transformation associated to the butterfly circuit with parameters $\theta_{\text{ran}}^{(i)}$ chosen randomly (we choose each parameter uniformly from $[0, 2\pi)$).

Figure 8 shows that the qWGAN is effective at learning mixed states of 4 qubits, though learning is clearly more challenging as the rank of the target density matrix increases. We recognize that the form of the generator (17) may not be well suited to optimization over mixed states. For example, it is often the case that different circuits in the generator optimize

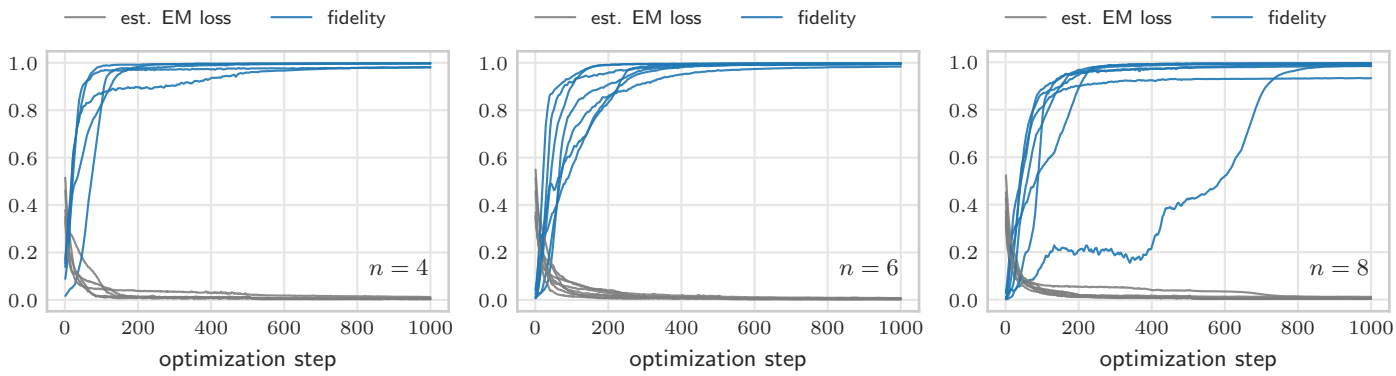


Figure 6: The student circuit is able to approximate well the state generated by the teacher circuit. Here, the target is constructed by randomly setting the parameters of a depth 2 mixing circuit (teacher circuit). The qWGAN, equipped with a generator circuit of depth 4, successfully learns the target state generated by the teacher circuit. For each plot, 5 simulations are performed.

to the same critical point in the loss landscape, thus outputting the same state. Future improvements to the design of generators can improve the results shown here.

J.5 QAOA learning

The quantum approximate optimization algorithm and its related extension to the quantum alternating operator ansatz, both given the acronym QAOA, are promising candidates for achieving quantum speedups in classical optimization problems [50–52]. Recent work has shown that QAOA is computationally universal [93] and potentially an effective algorithm in a wide range of quantum machine learning settings [11, 94, 138–141].

Here, we use a QAOA circuit as the generator for our qWGAN to learn the ground state of a simple translationally invariant Ising Hamiltonian cost function C:

$$C = B \sum_{i=1}^{N} \sigma_Z^{(i)} \sigma_Z^{(i+1)}, \tag{46}$$

where B is a constant assumed to be positive and $\sigma_Z^{(i)}$ is the Pauli Z operator acting on qubit i. Given the simple translationally invariant form of C, its ground state is spanned by the states $|01\rangle^{\otimes \frac{1}{2}n}$ and $|10\rangle^{\otimes \frac{1}{2}n}$.

For our experiments, we attempt to learn the ground state of C: $\frac{1}{\sqrt{2}}\left(|01\rangle^{\otimes \frac{1}{2}n} + |10\rangle^{\otimes \frac{1}{2}n}\right)$. We use a QAOA circuit which applies, repeating for a depth of L times, a mixing Hamiltonian $e^{-i\alpha_l H_{\text{mix}}}$ and the cost Hamiltonian $e^{-i\beta_l H_C}$ where $l \in \{1, \ldots, L\}$ indicates the layer of the QAOA circuit. In total, the circuit has $2L$ trainable parameters α_l and β_l (see Appendix K for details of circuit).

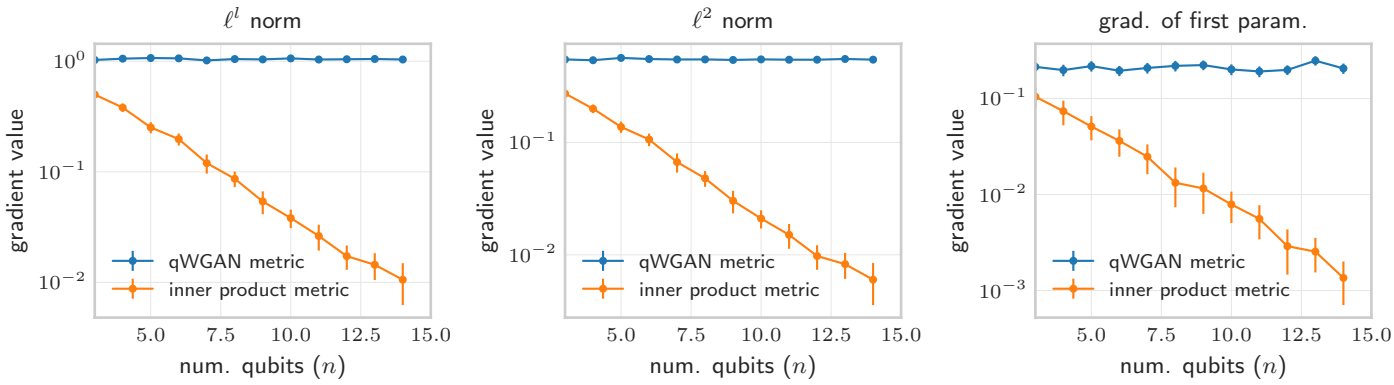


Figure 7: Comparison of gradients between the quantum EM loss metric and a conventional loss metric that is a function of the inner product. Here the average L1 norm (left), L2 norm (center), and the absolute value of the gradient of the first parameter (right) are shown. Decaying gradients observed in the inner product loss metric. In contrast, regardless of the number of qubits, the gradients of the qWGAN remain stable. Gradients are calculated at first step of optimization. L_1 norm and L_2 norm are divided by n and $\sqrt{n}$ respectively to normalize based on the number of parameters in the circuit. Results are averaged across 100 simulations for each data point

$$H_{\text{mix}} = \sum_{i=1}^{N} \sigma_X^{(i)} \qquad H_C = \sum_{i=1}^{N} \sigma_Z^{(i)} \sigma_Z^{(i+1)} \tag{47}$$

Figure 9 shows that our qWGAN is very effective at learning the ground state using the QAOA circuit as the generator. Convergence to the ground state is achieved within a few hundred steps of optimization.

Appendix K Circuits Used in Experiments

In all our experiments, the generators are parameterized circuits. The form of those circuits are listed below.

- GHZ circuit (subsection 5.1): circuit is shown in Figure 10. This circuit differs from that used in the toy model (Figure 1a) only in the first qubit. Here, three parameterized Pauli rotations are applied to the first qubit to allow for complete control over the relative phase of the first qubit.

- Mixing circuit (subsection 5.2): circuit is shown in Figure 11. This circuit is commonly used in prior literature to show the existence of barren plateaus in the loss landscape [20, 22]. This circuit contains alternating layers of parameterized Pauli Y rotations and pairwise Pauli Z-Z rotations.

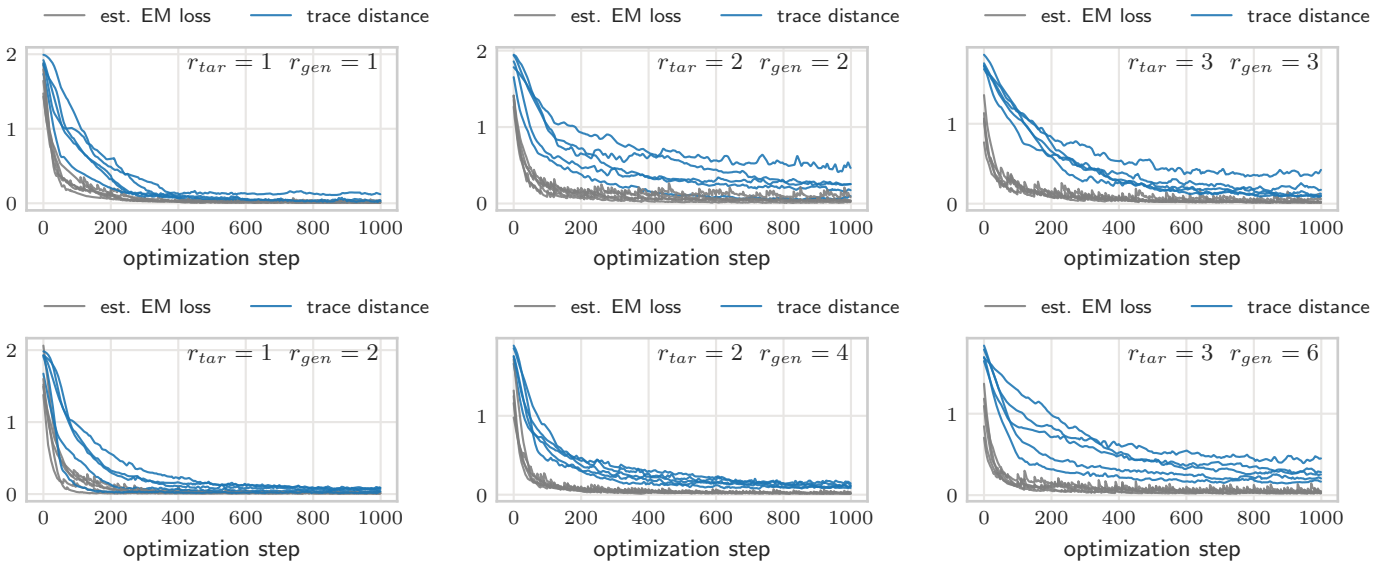


Figure 8: The qWGAN is also able to generate mixed states that approximate well their given target (also a mixed state) in both trace distance and quantum EM distance. Here, the generator circuit takes the form of a butterfly circuit of 4 qubits (see Appendix K), and the target is constructed by randomly setting the parameters of the generator circuit. The qWGAN aims to learn the target density matrix of rank r_{tar} with either $r_{\text{gen}} = r_{\text{tar}}$ or $r_{\text{gen}} = 2r_{\text{tar}}$ parameterized circuits of the same form. For each plot, 5 simulations are performed.

- Butterfly circuit (subsection J.4): The butterfly circuit takes the form of alternating layers of single qubit Pauli X rotations followed by controlled Pauli X rotations applied in the order of the butterfly pattern (Figure 12a). The form of the circuit for 4 qubits shown in Figure 12b.
- QAOA circuit (subsection J.5): general form of circuit is shown in Figure 13a consisting of alternating applications of a mixing Hamiltonian H_{mix} and cost Hamiltonian H_C. An initial layer of Hadamard gates is also included. At a given layer l, Trotterized time evolution circuits are used to apply H_{mix} and H_C for times α_l and β_l respectively [142]. The form of the circuit for 4 qubits and a single QAOA layer ($L = 1$) is shown in Figure 13b.

Appendix L Computational Details

All code used for this paper is available here: *redacted until publication*

Quantum circuit simulations were performed using PennyLane [142] with a backend of Tensorflow [143] or Pytorch [144]. Unless specified otherwise, the Adam optimizer is used

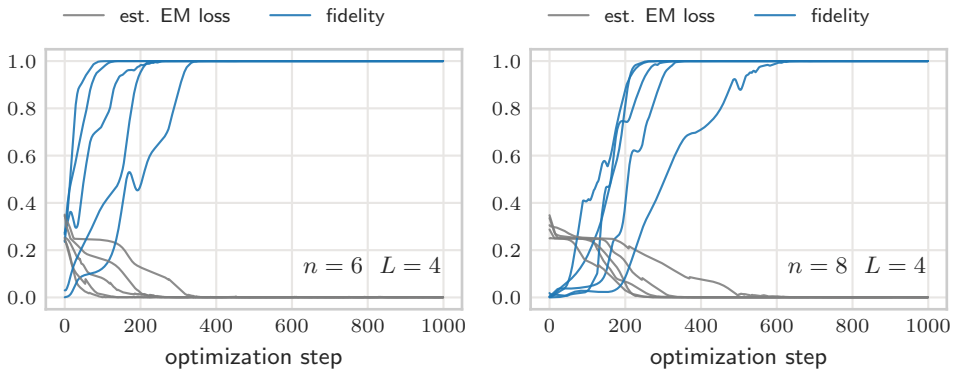


Figure 9: The qWGAN is effective at learning the ground state of a translationally invariant Ising Hamiltonian. Here, the generator is a QAOA circuit (see Appendix K) of depth $L = 4$. Estimated W_1 loss (quantum EM distance estimated by active operators) is also plotted in above chart, normalized by dividing by the number of qubits. For each plot, 5 simulations are performed.

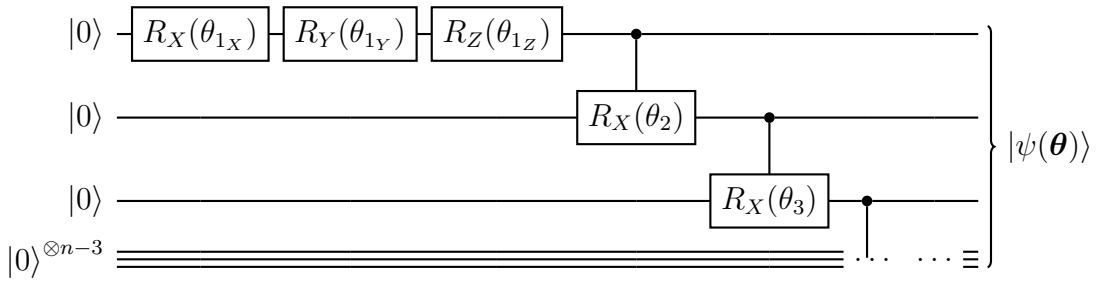


Figure 10: Circuit for generator in GHZ simulations (subsection 5.1).

for performing gradient-based updates on a generator [131]. The default Adam optimizer was set to a learning rate of 0.01. In some cases, learning was performed in two phases, first with a learning rate of 0.02 decreased to 0.007 for a second phase.

All parameters of the generator are initialized according to a standard normal distribution unless otherwise stated. In its default setting, we cycle the operators of the discriminator every ten optimization steps. When operators are cycled, a cycling threshold of $c = 0.8$ is used (see section 4.1). Discriminators are initialized with the set of 2-local Pauli operators.

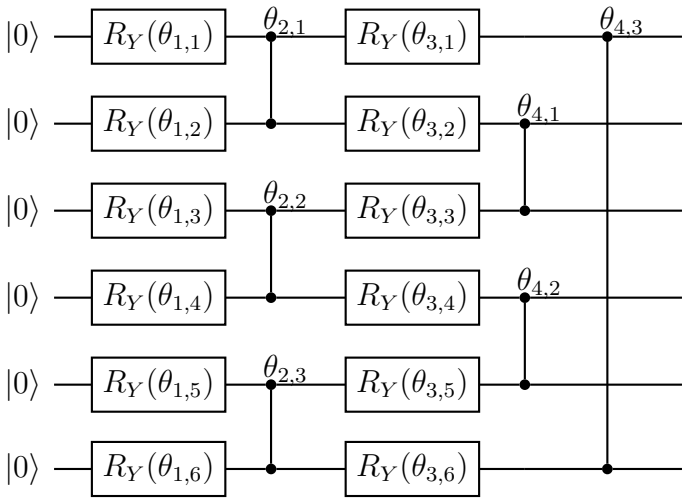
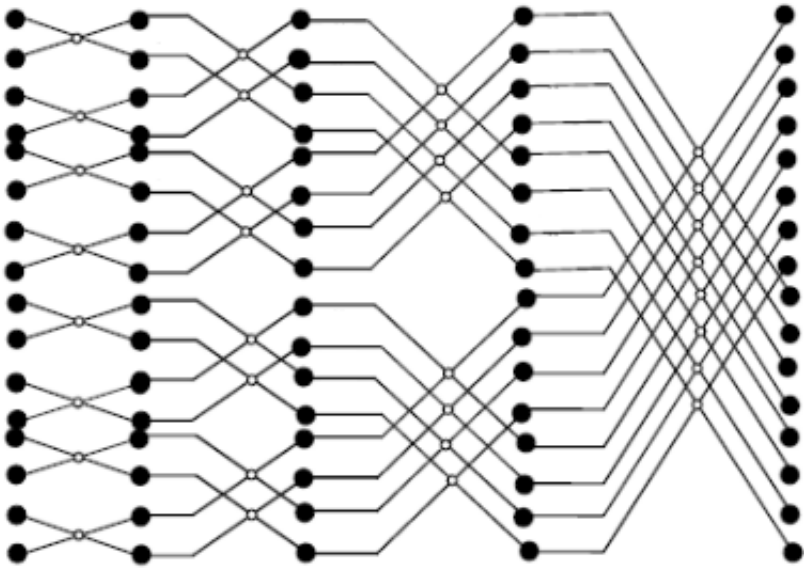
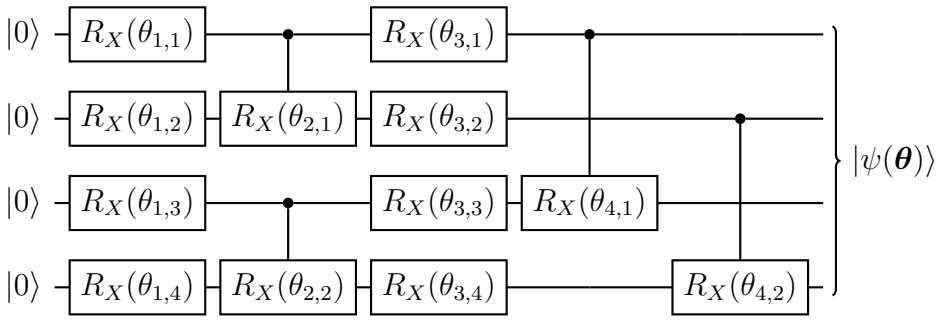


Figure 11: Single layer of mixing circuit used to perform learning and generate targets in subsection 5.1. This circuit consists of alternating layers of parameterized Pauli Y rotations and parameterized Pauli Z-Z rotations. The circuit above may be repeated to construct deeper circuits for simulations.

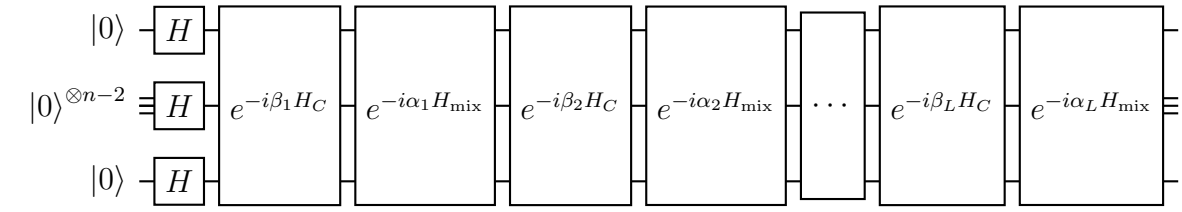


(a)

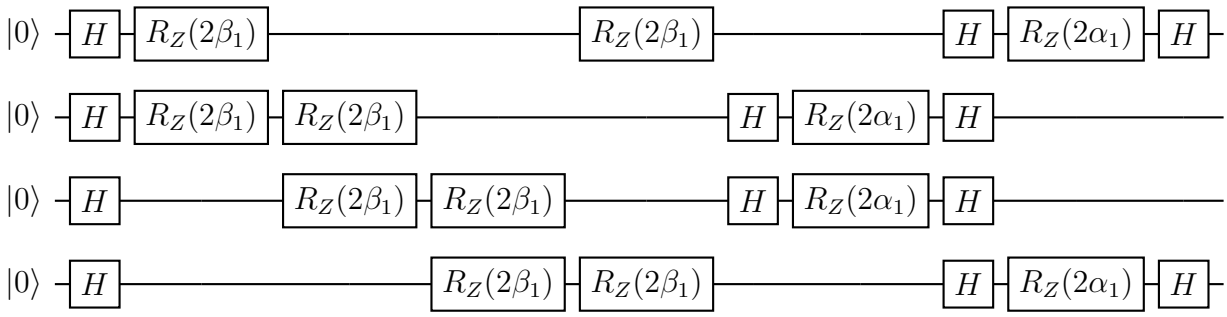


(b)

Figure 12: (a) Butterfly pattern of interactions, here shown for a system of 16 qubits. (b) Circuit for generator in butterfly circuit simulations (subsection J.4) here shown for 4 qubits.



(a)



(b)

Figure 13: (a) General form for QAOA circuit. (b) QAOA circuit for 4 qubits and $L = 1$.