Learning quantum data with the quantum earth mover's distance

Quantum Science and Technology

OPEN ACCESS

RECEIVED
16 May 2022

ACCEPTED FOR PUBLICATION
17 June 2022

PUBLISHED
4 July 2022

PAPER

Learning quantum data with the quantum earth mover's distance

Bobak Toussi Kiani[1,2,*], Giacomo De Palma[2,3,4,5], Milad Marvian[6], Zi-Wen Liu[7] and Seth Lloyd[2,3]

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[3] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[4] Scuola Normale Superiore, Pisa, Italy
[5] Department of Mathematics, University of Bologna, Bologna, Italy
[6] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, United States of America
[7] Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada
[*] Author to whom any correspondence should be addressed.

E-mail: bkiani@mit.edu

Keywords: quantum computation, artificial intelligence, earth mover's distance, Wasserstein distance, generative learning, quantum information

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. To overcome these obstacles, we consider here the recently proposed quantum earth mover's (EM) or Wasserstein-1 distance 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. We provide examples where our qWGAN is capable of learning a diverse set of quantum data with only resources polynomial in the number of qubits.

1. Introduction

A fundamental task in quantum machine learning is designing efficient algorithms for learning quantum states [1–10], transformations [10–17], and classical data stored as or generated by quantum states [18–20]. 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 an objective function (often a distance metric) which determines how close a generated object is to its target. This choice of 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 [21–23] and poor local minima [24–26] 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 [27]. 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 EM distance are often more stable and avoid issues with vanishing or exploding gradients [28–32] (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.

In this work, we study the quantum EM distance from an applied setting and make the following contributions. First, we overview the construction of the quantum EM distance and 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 operationalize the quantum EM distance, we devise a new heuristic method to approximate the quantum EM distance efficiently given copies of quantum states. In learning settings, 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 (GAN) [28] (see also [9]). Importantly, like its classical analog, our qWGAN employs an EM distance in its cost function. Numerical results show that our qWGAN is efficient at learning quantum data with shallow circuits in various settings. Finally, we discuss near term applications of our qWGAN for both classical and quantum problems.

2. Quantum distance metrics and quantum EM distance

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, GAN provide a neural network approach for learning a target probability distribution and generating new samples from the approximate distribution [28, 33]. 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 qGAN [7, 9, 26, 34]. 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 EM distance recently defined in [27] 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, 34]) 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.

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 [27] and also known as the 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 [28], D_{EM} possesses several properties that are desirable when learning quantum data.

The quantum EM distance of [27] 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 $\|\cdot\|_{\text{EM}}$ the corresponding norm, whose analytical expression can be found below. 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 (equal to the number of elements that differ between two strings), i.e., for any two strings of n bits x and y we have $D_{\text{EM}}(|x\rangle\langle x|, |y\rangle\langle y|) = h(x, y)$. 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 [27], 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

$$\|H\|_L = 2 \max_{i=1,\ldots,n} \min \left\{ \|H - \tilde{H}_{\bar{i}}\|_\infty : \tilde{H}_{\bar{i}} \in \mathcal{O}_n \text{ acts as identity on the } i \text{ th qubit} \right\}. \tag{3}$$

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 [27]. 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:

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

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.

Equivalently, the quantum EM distance can be also defined by its primal formulation [27, definition 6]:

$$D_{\text{EM}}(\rho, \sigma) = \frac{1}{2} \min \left\{ \sum_{i=1}^{n} \|X_i\|_1 : X_i \in \mathcal{O}_n,\ \text{Tr}_i X_i = 0 \quad \forall\ i = 1, \ldots, n,\ \sum_{i=1}^{n} X_i = \rho - \sigma \right\}. \tag{5}$$

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.

2.1. 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 [35]. 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.

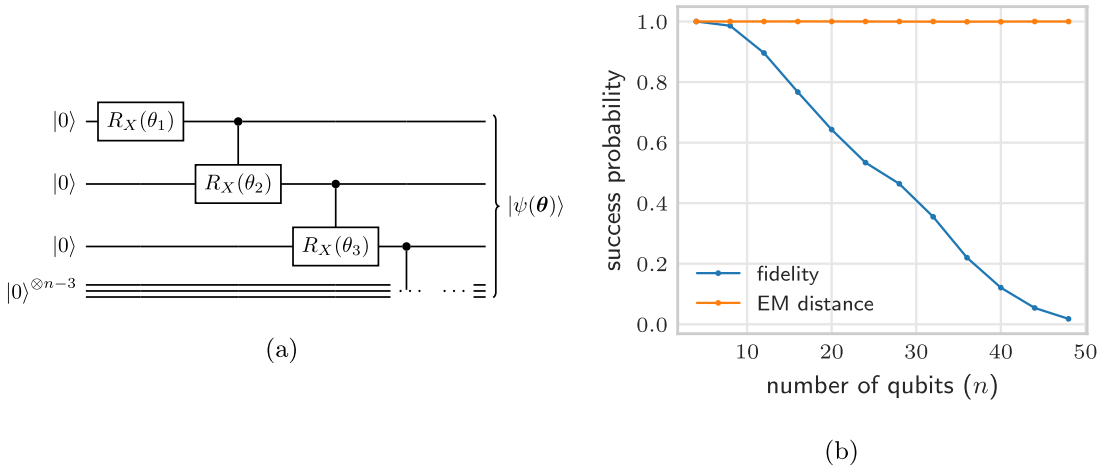

Figure 1. (a) Simple quantum circuit which can generate the GHZ state when parameter values are appropriately chosen. (b) In learning settings, this circuit presents challenges for loss metrics that are a function of the inner product between target and initial states. Simulations show that, with a loss 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 E for full details of experiments.

In this toy model, the task at hand is to learn the correct values of parameters in the circuit in figure 1(a) to generate the GHZ state of n qubits $|\mathrm{GHZ}_n\rangle = \left(|0_n\rangle + |1_n\rangle\right)/\sqrt{2}$. 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 $|\mathrm{GHZ}_n\rangle$. Given our circuit, F takes a simple form:

$$F = |\langle \mathrm{GHZ}_n|\psi(\boldsymbol{\theta})\rangle|^2 = \left(\frac{\cos(\theta_1)}{\sqrt{2}} + \frac{\prod_{i=1}^{n}\sin(\theta_i)}{\sqrt{2}}\right)^2. \tag{6}$$

The first problem associated with learning via F is that the loss landscape has undesirable local minima associated with the 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 [36]), 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.

$$\frac{\partial F}{\partial \theta_i} = \begin{cases} -\cos\,\theta_1\,\sin\,\theta_1 + \cos\,2\,\theta_1 \prod_{k=2}^{n}\,\sin\,\theta_k & i = 1 \\ +\sin\,\theta_1\,\cos\,\theta_1 \left(\prod_{k=2}^{n}\,\sin\,\theta_k\right)^2 & \\ \cos\,\theta_i \prod_{k\neq i}\sin(\theta_k)\left[\cos(\theta_1) + \prod_{k=1}^{n}\,\sin(\theta_k)\right] & i > 1 \end{cases}. \tag{7}$$

Notably, $\frac{\partial F}{\partial \theta_i} = O(\frac{1}{2^n})$ for $i > 1$, but $\frac{\partial F}{\partial \theta_i} = O(1)$ for $i = 1$. For large n, $\frac{\partial F}{\partial \theta_1} \approx -\cos\,\theta_1\,\sin\,\theta_1$ indicating that gradient optimizers will converge to a poor local minimum outputting the state $|0_n\rangle$ ($\theta_1 = 0 \bmod 2\pi$). In fact, since the loss function F (equation (6)) takes a simple form, gradient descent on the parameters can be efficiently performed classically, and results shown in figure 1(b) show that gradient descent converges to the undesirable local optimum associated with the $|0_n\rangle$ state even as more qubits are added (simulations stopped after 100 000 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 1(a), 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 intuitively explains this success and shows that the sequence of states $|\Psi_0\rangle = |0_n\rangle, |\Psi_1\rangle, \ldots, |\Psi_n\rangle = |\mathrm{GHZ}_n\rangle$ gets closer and closer to the target state $|\mathrm{GHZ}_n\rangle$, with a guaranteed improvement every two steps. The proof is in appendix F.

Proposition 1. *For any $k = 0, \ldots, n$, let $D_k = D_{\mathrm{EM}}\left(|\Psi_k\rangle\langle\Psi_k|, |\mathrm{GHZ}_n\rangle\langle\mathrm{GHZ}_n|\right)$. We have $n/2 \leqslant D_0 \leqslant (n+1)/2$, $D_n = 0$, and $(n-k)/2 \leqslant D_k \leqslant \left(n-k+\sqrt{2}\right)/2$ 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 1(b), optimization using the quantum EM distance is, in virtually all cases, successful at learning the GHZ state. In the simulations for figure 1(b), the quantum EM distance is efficiently estimated (lower bounded) 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.

$$\begin{aligned}\tilde{D}_{\mathrm{EM}} &= \max_{H_i} |\langle\psi(\boldsymbol{\theta})|H_i|\psi(\boldsymbol{\theta})\rangle - \langle\mathrm{GHZ}_n|H_i|\mathrm{GHZ}_n\rangle| \\ &\leqslant D_{\mathrm{EM}}\left(|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|, |\mathrm{GHZ}_n\rangle\langle\mathrm{GHZ}_n|\right), \end{aligned} \tag{8}$$

where $\tilde{D}_{\mathrm{EM}}$ 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 E for list of operators). Using $\tilde{D}_{\mathrm{EM}}$, we can successfully learn and generate the GHZ state regardless of the size of the system. Given the simplified form of our circuit, calculating $\tilde{D}_{\mathrm{EM}}$ can be efficiently performed using a classical computer and the methodology is detailed in appendix E. Interestingly, though the subset of Hamiltonians considered in calculating $\tilde{D}_{\mathrm{EM}}$ is significantly less than the total space of Hamiltonians available needed to exactly calculate D_{EM}, the simplified form of $\tilde{D}_{\mathrm{EM}}$ 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.

2.2. Properties of quantum EM distance in learning settings

For the EM distance over probability distributions, the classic work of [28] showed that the EM distance has a number of properties that confer advantages in learning settings over other distance metrics such as the total variational distance. Here, we show that our quantum EM distance offers corresponding analogous properties when learning in quantum settings. These properties provide intuitive explanations for why learning was so successful in the toy model analyzed earlier. First, the quantum EM distance is super-additive with respect to the tensor product [27, proposition 4]:

Proposition 2. *For any two quantum states ρ, σ of n qubits and any $k = 1, \ldots, n-1$,*

$$D_{\mathrm{EM}}(\rho, \sigma) \geqslant D_{\mathrm{EM}}(\rho_{1\ldots k}, \sigma_{1\ldots k}) + D_{\mathrm{EM}}(\rho_{k+1\ldots n}, \sigma_{k+1\ldots n}), \tag{9}$$

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.

In the case of product states where $\rho = \rho_{1\ldots k} \otimes \rho_{k+1\ldots n}$ and $\sigma = \sigma_{1\ldots k} \otimes \sigma_{k+1\ldots n}$, then the above is an equality:

$$D_{\mathrm{EM}}(\rho, \sigma) = D_{\mathrm{EM}}(\rho_{1\ldots k}, \sigma_{1\ldots k}) + D_{\mathrm{EM}}(\rho_{k+1\ldots n}, \sigma_{k+1\ldots n}), \tag{10}$$

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 $|\mathrm{GHZ}_2\rangle|1\rangle$, updating the state $|000\rangle$ to $|\mathrm{GHZ}_2\rangle|0\rangle$ 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. As an aside, super-additivity is relevant in noisy contexts as it implies that if noise only effects a small number of qubits, then the change in the EM distance is correspondingly bounded by the number of qubits on which the noise acts.

A second useful property of the quantum EM distance is that it recovers the classical EM 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 [27, proposition 6]:

Proposition 3. *Let p, q be probability distributions on $\{0,1\}^n$, and let*

$$\rho = \sum_{x\in\{0,1\}^n} p(x)\,|x\rangle\langle x|, \qquad \sigma = \sum_{y\in\{0,1\}^n} q(y)\,|y\rangle\langle y|. \tag{11}$$

Then, $D_{\mathrm{EM}}(\rho,\sigma) = D_{\mathrm{EM}}(p,q)$. In particular, the quantum EM distance between vectors of the canonical basis coincides with the Hamming distance: $D_{\mathrm{EM}}(|x\rangle\langle x|, |y\rangle\langle y|) = h(x,y)$ 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 [27, proposition 2]:

Proposition 4. *For any two quantum states ρ, σ,*

$$D_1(\rho,\sigma) \leqslant D_{\mathrm{EM}}(\rho,\sigma) \leqslant n\,D_1(\rho,\sigma). \tag{12}$$

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.

2.3. EM distance evaluation

As in the classical Wasserstein GAN [28], approximations to the EM distance are required to construct learning algorithms that have efficient runtimes. Note, the quantum EM distance can be exactly evaluated using algorithms for semidefinite programs [37, 38] which run in time polynomial in the dimension of the quantum state and the number of constraints. Such an exact approach would require algorithmic runtimes that are exponential in the number of qubits and furthermore, do not 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, we construct a parameterized family of functions which preserve a quantum Lipschitz constraint upon optimization. Let

$$H = \sum_{\mathcal{I}\subseteq\{1,\ldots,n\}} H_{\mathcal{I}}, \tag{13}$$

where each $H_{\mathcal{I}}$ acts non-trivially only on the qubits in the corresponding set $\mathcal{I}$. Then, proposition 10 of [27] provides an upper bound to the quantum Lipschitz constant of a Hamiltonian in terms of its local structure.

$$\|H\|_L \leqslant 2 \max_{i=1,\ldots,n} \left\| \sum_{i\in\mathcal{I}\subseteq\{1,\ldots,n\}} H_{\mathcal{I}} \right\|_\infty, \tag{14}$$

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. Intuitively, since the Lipschitz constant bounds the change in a Hamiltonian induced by changes to a single qubit, the bound above can be viewed as a bound on the maximum singular value of nontrivial operators thus also bounding the corresponding Lipschitz constant. A natural choice for the operators $H_{\mathcal{I}}$ are a subset of the Pauli operators which we explore in our construction of a qGAN next.

3. 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 $D_{\mathrm{EM}}(G(\theta), \rho_{\mathrm{tar}})$. 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.

3.1. 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:

$$H(W) = \sum_{P_1, \ldots, P_n \in \{I,X,Y,Z\}} w_{P_1 \ldots P_n} \, \sigma_{P_1}^{(1)} \otimes \sigma_{P_2}^{(2)} \otimes \ldots \otimes \sigma_{P_n}^{(n)}, \tag{15}$$

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. Finding the exact Lipschitz constant of the Hamiltonian (15) can be computationally expensive. However, noting that Pauli operators have infinity norm of 1 and applying the triangle equality, (14) provides an easily computable upper bound to $\|H(W)\|_L$, which we denote with $\|H(W)\|_{\tilde{L}}$:

$$\|H(W)\|_{\tilde{L}} = 2 \max_{i=1, \ldots, n} \sum_{P_1, \ldots, P_n \in \{I,X,Y,Z\}: P_i \neq I} |w_{P_1 \ldots P_n}| \geqslant \|H(W)\|_L. \tag{16}$$

The Hamiltonian (15) 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 (i.e., acting on k qubits and not necessarily geometrically local) Pauli operators as the discriminator. We denote with $\mathcal{O}_n^{(k)}$ the linear span of such operators. For example, the most general element of $\mathcal{O}_n^{(2)}$ is

$$H(W) = w_I + \sum_{i=1}^{n} \sum_{P \in \{X,Y,Z\}} w_P^{(i)} \, \sigma_P^{(i)} + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{P,\, Q \in \{X,Y,Z\}} w_{P,Q}^{(i,j)} \, \sigma_P^{(i)} \otimes \sigma_Q^{(j)}, \tag{17}$$

where each w_I, $w_P^{(i)}$ and $w_{P,Q}^{(i,j)}$ 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.

We can now define an approximated EM distance by restricting the optimization in (4) to k-local Hamiltonians and replacing the exact Lipschitz constant with the approximated Lipschitz constant (16):

$$D_{\mathrm{EM}}^{(k)}(\rho, \sigma) = \max \left\{ \mathrm{Tr}\left[(\rho - \sigma) H\right] : H \in \mathcal{O}_n^{(k)}, \ \|H\|_{\tilde{L}} \leqslant 1 \right\}. \tag{18}$$

The approximated quantum EM distance is increasing with respect to k and provides a lower bound to the exact quantum EM distance, i.e., for any two quantum states ρ and σ,

$$D_{\mathrm{EM}}^{(1)}(\rho, \sigma) \leqslant D_{\mathrm{EM}}^{(2)}(\rho, \sigma) \ldots \leqslant D_{\mathrm{EM}}^{(n)}(\rho, \sigma) \leqslant D_{\mathrm{EM}}(\rho, \sigma). \tag{19}$$

The order k of Pauli operators can be tuned to the complexity of a given problem. To learn truly random states, access to the full set of Hamiltonian operators is needed, but in many cases, especially when learning data generated by shallow circuits or data that can be discriminated via reduced density matrices, learning using only lower order Pauli operators can be effective.

The approximated EM distance (18) can be computed with the following linear program, which can be efficiently solved. 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:

$$\begin{aligned} &\text{maximize} \sum_{j=1}^{|W|} c_j w_j \\ &\text{subject to} \sum_{j: i \in \mathcal{I}_j} |w_j| \leqslant 1, \quad i = 1, \ldots, n \end{aligned} \tag{20}$$

where c_j is the trace of the product between the jth Pauli string and $G(\theta) - \rho_{\mathrm{tar}}$. i.e., assuming w_j is associated to Pauli string $\sigma_{P_a}^{(a)} \sigma_{P_b}^{(b)} \ldots \sigma_{P_k}^{(k)}$, then $c_j = \mathrm{Tr}\left[(G - \rho_{\mathrm{tar}}) \sigma_{P_a}^{(a)} \, \sigma_{P_b}^{(b)} \ldots \sigma_{P_k}^{(k)}\right]$. 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 (20) 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 [39] (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 $n_{\text{active}} \leqslant n$ operators with non-zero weights, called active operators, constructing a Hamiltonian $H_{\max}$ which is passed onto the generator for optimization:

$$H_{\max} = \sum_{i=0}^{n_{\text{active}}} w_i^* H_i^*, \tag{21}$$

where w_i^* and H_i^* are the weights and active operators respectively. As an example, we show in appendix H that for single qubit states, the linear program outputs an optimal Hamiltonian composed of single qubit Pauli terms chosen by selecting the single qubit Pauli term which has the greatest contribution on each qubit (thus $n_{\text{active}} = n$ in this setting). Since $n_{\text{active}} \leqslant n$, gradient updates on the generator which aims to minimize the expectation of $H_{\max}$ can be performed. However, this efficiency comes at the cost of potentially missing certain active operators in the optimal Hamiltonian. Equation (21) is naturally biased towards including low order operators in its set of active operators. Especially for problems where the optimization needs to access higher order operators, this methodology may fail to perform effectively.

Appendix H compares values of the estimated distance $D_{\text{EM}}^{(2)}$ with the actual distance D_{EM} for random states generated by shallow circuits showing that there is a correlation between the two measures, although the approximation may introduce an unwanted bias. As we show in appendix G, 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 (20) contains only these terms whenever the coefficients c_j associated to single Pauli operators are all $\Omega(1)$. Furthermore, since operators only act on few qubits, efficient algorithms from [40, 41] can be applied to calculate the expectations of $|W|$ Pauli operators in (17) using $O(\log|W|)$ copies of the generated and target states. If a large number of higher order Pauli terms are included in the decomposition above, such logarithmic dependence may no longer hold.

We stress that the approximation employed here is the central limitation in successfully applying our qWGAN model. Improving this approximation is crucial to expanding the scope of application of our qWGAN beyond the assorted examples provided in section 4. Generally, there are two paths by which one can improve the approximation. The first path consists of identifying a more optimal relaxation to the semidefinite program in comparison to the linear relaxation of equation (20). Existing quantum algorithms for solving semidefinite programs [37, 38] may help in realizing this goal. The second path consists of finding methods to more optimally choose the subset of Hamiltonians or Pauli operators included in the approximation. In some cases, higher order Pauli operators are needed to distinguish states and avoiding these altogether may produce sub-optimal results. In light of this, we describe below one technique that empirically helps in finding a good subset of Pauli operators.

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 operators (here, we choose the new operators uniformly randomly from the set of all Pauli 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 $P(\min_i c_i^*)$ (i.e., minimum taken over all active operators) where $0 < P \leqslant 1$. When an operator is cycled out, a random Pauli operator is then included in the discriminator's set of operators including potentially Pauli operators that were removed in earlier cycles. Cycling operators may, of course, be detrimental if operators are cycled out that end up being useful during later phases of training. Nevertheless, in our experiments, we often find that the amount of cycling can serve as another tunable hyperparameter for improving the performance of our qWGAN.

3.2. 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 [42–44] and transformer networks optimized for text analysis [45–47]—the form of the generator in our quantum

Algorithm 1. qWGAN with quantum earth mover's distance.

Require: initial discriminator operators: $H_i^{[0]}$ ▷ e.g., set of two-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: $c_i \leftarrow \mathrm{Tr}\,[H_i(G(\theta) - \rho_{\mathrm{tar}})]$
  find w_i^*, H_i^* (linear program, equation (20)) ▷ $H_{\max} = \sum_i w_i^* H_i^*$
  optional: cycle operators
  ▷ *generator optimization:*
  find gradients g_p, g_θ of $\mathrm{Tr}[G(\theta)H_{\max}]$ ▷ See appendix I
  perform gradient update on θ and p ▷ e.g., $\theta \leftarrow \theta - \alpha g_\theta$

algorithm can and should be customized to the specific problem setting. Many options exist for constructing a generator including parameterized quantum circuits [48, 49] and quantum neural networks [50–53]. 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 $|\psi_0\rangle\langle\psi_0|$ 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 $\{(p_1, U_1), \ldots, (p_r, U_r)\}$:

$$G(\theta) = \sum_{i=1}^{r} p_i U_i |\psi_0\rangle\langle\psi_0| U_i^\dagger, \tag{22}$$

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 4 for examples).

3.3. 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 [28]. 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 (20) to maximize the quantum EM distance D_{EM} in equation (4). Then, a gradient update is performed on the parameters of the generator θ.

3.4. Properties of the gradient

As stated earlier, one can calculate the gradients with respect to the parameters of the unitary operator implemented by the circuit (step 4 in algorithm 1) via the parameter shift rule [54]. As an example, let $G(\theta)$ be generated by the quantum circuit that implements the unitary operator

$$U(\theta) = \prod_k U_k \, \mathrm{e}^{-\mathrm{i}\theta_k P_k}, \tag{23}$$

where each U_k is a unitary operator that does not contain any parameter and each P_k is a generalized Pauli operator. Then, given an optimal Hamiltonian $H_{\max}$, one can calculate the gradient as follows for a given parameter θ_k:

$$\frac{\partial}{\partial\theta_k}\mathrm{Tr}[G(\theta)H_{\max}] = \mathrm{Tr}[G(\theta^+)H_{\max}] - \mathrm{Tr}[G(\theta^-)H_{\max}], \tag{24}$$

where θ^+ and θ^- are the values of the parameters shifted by $\frac{\pi}{4}$ in either direction for the entry corresponding to θ_k [54]. This parameter-shift rule has the benefit that the equation for the gradient is in fact exact (not an approximation). Gradients for each individual parameter must be calculated using separate circuit evaluations.

Prior work has shown that local cost functions avoid barren plateaus up to poly-logarithmic depth in the circuit [23]. Due to the super-additivity property of the quantum EM distance and the construction of the linear relaxation in (20), the optimal Hamiltonian is heavily biased towards local terms and thus fits into

this regime. We formalize this below by showing that high order Pauli strings acting nontrivially on k qubits must have magnitude greater than k times the smallest single qubit Pauli contribution to the optimal Hamiltonian $H_{\max}$.

Proposition 5. *Let $w^* : \{I, X, Y, Z\}^n \to \mathbb{R}$ be the set of parameters that achieve the maximum in (20), and let*

$$a = \min_{i=1,\ldots,n} \max_{P=X,Y,Z} \left| \mathrm{Tr}\left[(G(\theta) - \rho_{tar})\, \sigma_P^{(i)} \right] \right|. \tag{25}$$

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

$$|c_{P_1 \ldots P_n}| < a\, |\{i = 1, \ldots, n : P_i \neq I\}|. \tag{26}$$

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

The proof of the above is deferred to appendix G. As a corollary of the above, any global k-qubit Pauli term that appears in the optimal Hamiltonian must exceed the sums of the maximum single qubit Pauli terms on the k qubits on which it acts. Thus, the optimal Hamiltonian $H_{\max}$ will be at most $2/a$-local which is constant when the expectation of single qubit Paulis is $\Omega(1)$. In these settings, the qWGAN will avoid barren plateaus whenever the generator has depth that does not grow faster than logarithmically in the number of qubits (see theorem 2 of [23]).

4. Experiments

Our qWGAN can efficiently learn quantum data of various forms. Here, we apply the qWGAN directly to the toy model of subsection 2.1 and also consider a more general scenario where the qWGAN learns states generated by a mixing circuit previously known to suffer from barren plateaus [21, 23, 55]. 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) [56–58] and one where the qWGAN is tasked with learning mixed states. The Adam optimizer with a default learning rate of 0.01 was used to train the qWGAN [59]. Details on the structure of the quantum circuits and on how the simulations were performed are provided in appendices K and L, respectively.

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 2.1, 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 for 10 random simulations, our qWGAN efficiently generates the $n = 8$ qubit GHZ state in around 500 steps of optimization—results for circuits of different size are also consistent with this analysis and detailed in appendix J. The generator circuit has $n + 2$ parameters and depth n. Simulations are repeated 50 times across random initializations. The discriminator starts with access to $k = 2$ local Pauli operators, cycling out 'bad' operators 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.

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 3(a). This circuit is a generic mixing circuit also studied in [21, 23, 55] 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 3(a) 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 3(b) 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

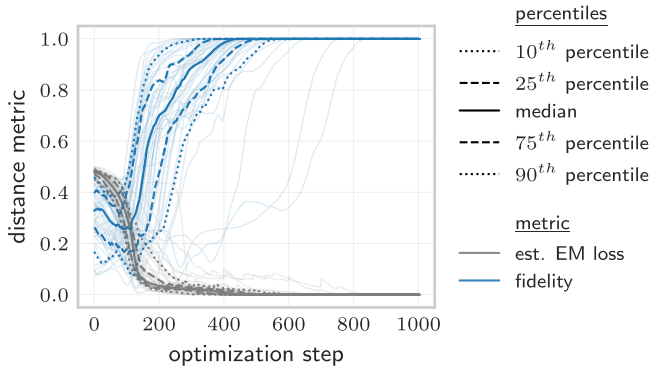

Figure 2. The qWGAN consistently generates the eight qubit GHZ state. Estimated EM loss (quantum EM distance estimated by active operators) is plotted in grey alongside the fidelity in blue. EM distance is normalized to a maximum of one by dividing by the number of qubits. Percentiles are calculated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

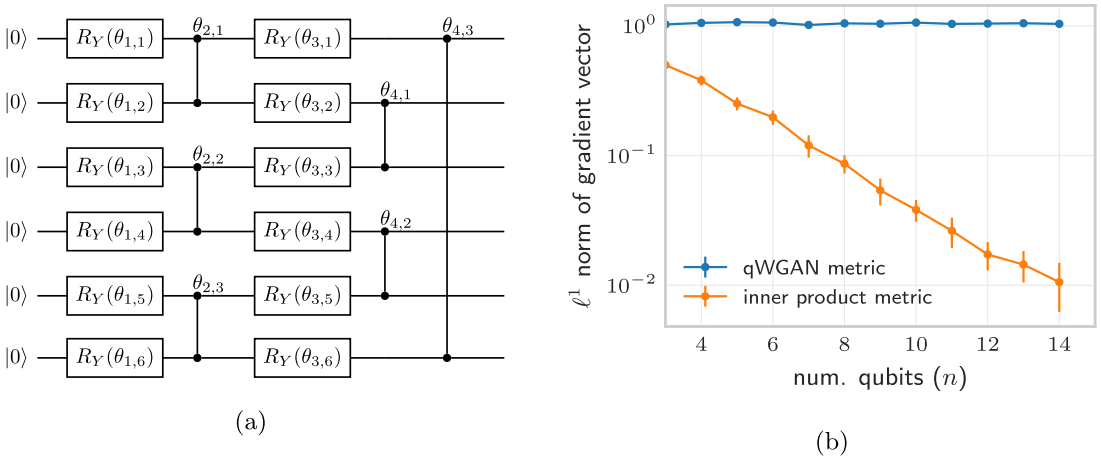

Figure 3. (a) Single layer of mixing circuit consisting of alternating layers of parameterized Pauli Y rotations and parameterized Pauli Z–Z rotations applied to pairwise qubits. Here, the form of the circuit is shown for six qubits. (b) Learning using two layers of mixing circuit (shallow, constant 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.

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 including the Wasserstein semi-metric proposed in [9].

Furthermore, as shown in figure 4, the qWGAN successfully learns the states constructed by $n = 8$ qubit teacher circuits using student circuits of depth 4. This circuit has 96 trainable parameters. Simulations are repeated 50 times across random initializations in figure 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. We find that, in general, learning in the teacher–student setting is best achieved when the student circuit is deeper than the teacher circuit. Furthermore, due to the approximations made in calculating the quantum EM distance, we find that our algorithm struggles to learn especially deep eight-qubit teacher circuits which are four layers or more in depth. For these deeper target circuits, we suspect that higher order Paulis are needed to efficiently estimate the quantum EM distance, and further improvements to the optimization must be made to incorporate these higher order Paulis in the estimation procedure. Additional simulations for different circuit sizes are shown in appendix J.

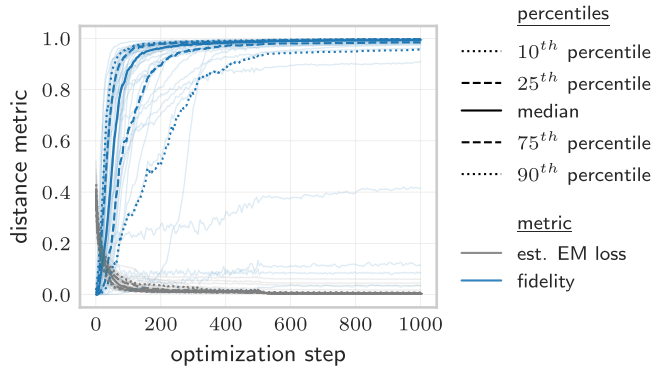

Figure 4. The student circuit is able to approximate well the state generated by the teacher circuit. Here, the target is constructed by randomly setting the parameters of a depth 2 mixing circuit (teacher circuit). The qWGAN, equipped with a depth 4 generator circuit, successfully learns the target state by optimizing over the quantum EM distance. Estimated EM loss (quantum EM distance estimated by active operators) is plotted in grey alongside the fidelity in blue. 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.

5. 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 can potentially confer advantages in machine learning settings, especially when optimization is performed over local gates in shallow circuits. 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 [28], we show that our qWGAN can potentially improve convergence and stability in learning quantum data. Nevertheless, the qWGAN struggles in learning more 'random' data or data generated from deep circuits. These challenges stem from two approximations made in the learning procedure. First, to ensure runtime efficiency, the discriminator was restricted to measuring a subset of local Pauli operators. Second, the optimization statement to calculate the quantum EM distance was relaxed into a linear program to ensure it can be efficiently calculated classically. Though this estimated distance is well correlated to the true distance when learning certain structured states like the GHZ state or states generated by shallow circuits, it failed to tightly bound the quantum EM distance in more challenging settings. Looking forward, improving the bounds given by relaxations of the quantum EM distance can potentially allow for application of our qWGAN in these more challenging settings.

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 [60, 61]. For unsupervised learning, the qWGAN provides a framework and approach to quantum circuit compression, data encoding, and sampling [62–64]. For quantum error correction, one can use a qWGAN to develop new techniques for constructing quantum error codes or assisting error correction procedures [65–68].

Author contributions

BTK developed the structure of the qWGAN algorithm and performed simulations under the supervision of SL. All authors contributed to the design of the qWGAN algorithm. The EM distance was motivated by prior work from GDP, MM, and SL. BTK and GDP wrote the manuscript. GDP developed mathematical proofs in the appendix. All authors reviewed the results.

Acknowledgment

This work was funded by AFOSR, ARO under the Blue Sky Initiative, DARPA, and NSF. MM is partially supported by the NSF Grant No. CCF-1954960 and DARP' RQMLS program. GDP is a member of the 'Gruppo Nazionale per la Fisica Matematica (GNFM)' of the 'Istituto Nazionale di Alta Matematica 'Francesco Severi' (INdAM)'.

Data availability statement

No new data were created or analysed in this study.

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 [21, 69, 70], cost functions with global observables [23], and noisy circuits [22]. Furthermore, research in quantum control theory has identified the presence of traps when the control landscape is constrained [24, 25, 71].

Various attempts have been made to potentially avoid these roadblocks. These include methods for initializing circuit parameters [72, 73], algorithms for layer-wise training [36], and choosing ansatzes that can avoid decaying gradients [74, 75]. Despite these efforts, prior work [23] has shown that barren plateaus are unavoidable unless cost functions are appropriately chosen to be local. Thus more pertinent to our work are prior studies that have specifically designed loss metrics or gradient-based algorithms that help avoid issues with barren plateaus. This includes reference [23], which considers local operator loss functions, but not necessarily distance metrics. The quantum EM distance, when used as a cost function, is biased towards local operators, gaining the benefits of using local cost functions outlined in [23]. Beyond direct changes to the cost functions, reference [55] includes second order derivatives in optimizing the loss function to better navigate flat loss landscapes, and reference [76] constructs an algorithm to optimize over the Fubini-study metric tensor. Though not analyzed here, these methods can be used in tandem with our qWGAN to improve optimization of circuit parameters and increase rates of convergence.

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 [18–20]. Reference [18] 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 GAN, a classical algorithm first introduced in [33]. Most relevant to the current work, reference [28] constructed the first classical Wasserstein GAN employing an EM distance. Later work improved upon the stability and training of the original Wasserstein GAN by, for example, constructing improved discriminators and generators [32, 77, 78], progressively adding layers during training [79], and employing various regularization techniques [80–82]. In the classical literature, GANs have been extensively used in many real-world applications [83–87].

In the quantum setting, quantum GANs were first proposed by references [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, 34, 88–90]. Hybrid classical-quantum GANs, fully classical in the discriminator, generator, and/or loss function, were proposed in references [91–94]. Reference [9] proposed a version of a quantum Wasserstein GAN (qWGAN), though the employed EM distance is unitarily invariant (see appendix D for details). Reference [95] 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 [34, 88, 96, 97].

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 [97, 98], chemistry [96], and post-processing quantum outputs [99–101]. 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 [58, 102, 103], the variational quantum eigensolver [104, 105], and quantum GANs discussed earlier.

Appendix B. The classical earth mover's distance

The classical EM distance, also called Monge–Kantorovich distance, is a distance between probability distributions on a metric space which dates back to Monge [106] 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 [107], 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 [31, 108–110].

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 [111].

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 $\mathrm{Tr}\,[\rho H]$.

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 $\mathrm{Tr}\,[\rho_1 H] = \mathrm{Tr}\left[\rho\left(H \otimes \mathbb{I}_{N_2}\right)\right]$ 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 ith qubit if it is equal to a 2×2 matrix acting on the ith 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).

$$\sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{27}$$

$$\sigma_Y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \tag{28}$$

$$\sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{29}$$

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:

$$R_P(t) = \mathrm{e}^{-\mathrm{i}t\sigma_P/2} = \cos\left(\frac{t}{2}\right) I - i \sin\left(\frac{t}{2}\right) \sigma_P \tag{30}$$

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 [112].

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 [113–118] 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 [117], and has led to determine the rate of convergence of the quantum Ornstein–Uhlenbeck semigroup [114, 119]. Exploiting their quantum differential structure, references [115, 116, 120] also define a quantum generalization of the Lipschitz constant and of the EM distance. Alternative definitions of quantum EM distances based on a quantum differential structure are proposed in references [121–124]. References [125–127] propose quantum EM distances based on a distance between the vectors of the canonical basis.

Another line of research by Golse, Mouhot, Paul and Caglioti [128–133] 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 $\mathrm{Tr}\,[\Pi\, C]$, 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. References [128–133] 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:

$$C = (Q_1 - Q_2)^2 + (P_1 - P_2)^2, \tag{31}$$

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. Reference [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, reference [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.

Reference [134] 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 [135]. The problem of defining a quantum EM distance through quantum couplings has been explored in reference [136].

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 references [137–139].

Appendix E. Toy model details

Our toy model (subsection 2.1) 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 $\boldsymbol{\theta}$.

$$F = |\langle \mathrm{GHZ}_n|\psi(\boldsymbol{\theta})\rangle|^2 = \left(\frac{\cos(\theta_1)}{\sqrt{2}} + \frac{\prod_{i=1}^{n} \sin(\theta_i)}{\sqrt{2}}\right)^2 \tag{32}$$

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 [59]. For each simulation, 100 000 steps of optimization were performed before stopping. Learning is considered successful if $1 - F < 0.02$.

In our toy model, to efficiently approximate the quantum EM distance, we construct a loss function $\tilde{D}_{\mathrm{EM}} \approx D_{\mathrm{EM}}\left(|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|, |\mathrm{GHZ}_n\rangle\langle\mathrm{GHZ}_n|\right)$ 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.

$$\begin{aligned}|\psi(\boldsymbol{\theta})\rangle &= \cos\,\theta_1|0_n\rangle + i\sin\,\theta_1\cos\,\theta_2|1\rangle|0_{n-1}\rangle - \sin\,\theta_1\sin\,\theta_2\cos\,\theta_3|1_2\rangle|0_{n-2}\rangle + \ldots \\ &= \cos\,\theta_1|0_n\rangle + \sum_{k=1}^{n-1} i^k \left[\prod_{j=1}^{k}\sin\,\theta_j\right]\cos\,\theta_{k+1}|1_k\rangle|0_{n-k}\rangle + i^n\left[\prod_{j=1}^{n}\sin\,\theta_j\right]|1_n\rangle\end{aligned} \tag{33}$$

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

$$|\psi(\boldsymbol{\theta})\rangle = \begin{bmatrix} \cos\,\theta_1 \\ i\sin\,\theta_1\cos\,\theta_2 \\ \vdots \\ i^n\prod_{j=1}^{n}\sin\,\theta_j \end{bmatrix}, \tag{34}$$

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

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

$$P_i \in \{\sigma_Z^{(1)},\ \sigma_Z^{(2)},\ldots,\ \sigma_Z^{(n)},\sigma_Y^{(1)},\ \sigma_X^{(1)}\otimes\sigma_X^{(2)},\ \sigma_X^{(1)}\otimes\sigma_X^{(2)}\otimes\sigma_Y^{(3)},\ldots,\ \sigma_X^{(1)}\otimes\sigma_X^{(2)}\otimes\ldots\otimes\sigma_X^{(n)}\}, \tag{35}$$

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 subsection 2.3, 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, $\tilde{D}_{\text{EM}}$ can be calculated as the maximum amongst the following $n+1$ parameterized sums of expectations of operators:

$$\begin{aligned}\tilde{D}_{\text{EM}} = \max\Big\{&\Big|\mathbb{E}[\sigma_Z^{(1)}]\Big| + \Big|\mathbb{E}[\sigma_Z^{(2)}]\Big| + \cdots + \Big|\mathbb{E}[\sigma_Z^{(n)}]\Big|, \Big|\mathbb{E}[\sigma_Y^{(1)}]\Big| + \Big|\mathbb{E}[\sigma_Z^{(2)}]\Big| + \cdots + \Big|\mathbb{E}[\sigma_Z^{(n)}]\Big|,\\ &\Big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_Y^{(2)}]\Big| + \Big|\mathbb{E}[\sigma_Z^{(3)}]\Big| + \cdots + \Big|\mathbb{E}[\sigma_Z^{(n)}]\Big|,\\ &\ldots,\\ &\Big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \ldots \otimes \sigma_X^{n}]\Big|\Big\},\end{aligned} \tag{36}$$

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:

$$\begin{aligned}\tilde{D}_{\text{EM}} = \max\Big\{&\Big|\mathbb{E}[\sigma_Z^{(1)}]\Big| + \Big|\mathbb{E}[\sigma_Z^{(2)}]\Big| + \cdots + \Big|\mathbb{E}[\sigma_Z^{(n)}]\Big|,\\ &\Big|\mathbb{E}[\sigma_Y^{(1)}]\Big|,\\ &\Big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_Y^{(2)}]\Big|,\\ &\ldots,\\ &\Big|\mathbb{E}[\sigma_X^{(1)} \otimes \sigma_X^{(2)} \otimes \ldots \otimes \sigma_X^{n}]\Big|\Big\}.\end{aligned} \tag{37}$$

Using the equation for $\tilde{D}_{\text{EM}}$ 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 [59]. Only up to 10 000 steps of optimization were performed since convergence was almost always achieved within about 1000 steps. Learning is considered successful if $|\langle \text{GHZ}_n|\psi(\boldsymbol{\theta})\rangle|^2 > 0.98$ after the optimization. Success was achieved in virtually all instances when using $\tilde{D}_{\text{EM}}$.

Appendix F. Proof of proposition 1

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

$$\Delta_k = |\Psi_k\rangle\langle\Psi_k| - |\text{GHZ}_n\rangle\langle\text{GHZ}_n|, \tag{38}$$

and let $\mathcal{D}_1$ be the completely dephasing channel acting on the first qubit. From [27, proposition 3], the quantum EM distance is contractive with respect to a quantum channel acting on a single qubit. We then have on the one hand

$$\|\Delta_k\|_{\text{EM}} \geqslant \|\mathcal{D}_1(\Delta_k)\|_{\text{EM}} = \left\||1_k\rangle\langle 1_k| \otimes \frac{|0_{n-k}\rangle\langle 0_{n-k}| - |1_{n-k}\rangle\langle 1_{n-k}|}{2}\right\|_{\text{EM}} = \frac{n-k}{2}. \tag{39}$$

On the other hand, we have

$$\begin{aligned}\|\Delta_k\|_{\text{EM}} &\leqslant \|\mathcal{D}_1(\Delta_k)\|_{\text{EM}} + \|\Delta_k - \mathcal{D}_1(\Delta_k)\|_{\text{EM}} = \frac{n-k}{2} + \frac{1}{2}\|\Delta_k - \mathcal{D}_1(\Delta_k)\|_1\\ &= \frac{n-k}{2} + \frac{1}{2}\begin{cases} 1 & k = 0\\ \sqrt{2} & k = 1, \ldots, n-1\end{cases},\end{aligned} \tag{40}$$

where the first equality follows from [27, proposition 2], stating that $\|X\|_{\text{EM}} = \|X\|_1/2$ for any $X \in \mathcal{O}_n$ with $\text{Tr}_1 X = 0$. The claim follows.

Appendix G. Bias towards local operators

Consider the maximization problem (20) copied below:

$$\begin{aligned} &\text{maximize} \sum_{j=1}^{|W|} c_j w_j \\ &\text{subject to} \sum_{j:i\in\mathcal{I}_j} |w_j| \leqslant 1, \quad i = 1, \ldots, n \end{aligned}$$

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

Proposition 6. *Let $w^* : \{I, X, Y, Z\}^n \to \mathbb{R}$ be the set of parameters that achieve the maximum in (20), and let*

$$a = \min_{i=1,\ \ldots,\ n} \max_{P=X,Y,Z} \left| \mathrm{Tr}\left[(G - \rho_{tar})\, \sigma_P^{(i)} \right] \right|. \tag{41}$$

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

$$|c_{P_1 \ldots P_n}| < a\, |\{i = 1,\ \ldots,\ n : P_i \neq I\}|. \tag{42}$$

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 (20) is a linear program with dual

$$\min_{z \in \mathbb{R}^n_{\geqslant 0}} \sum_{i=1}^{n} z_i \quad : \quad |c_{P_1 \ldots P_n}| \leqslant \sum_{i \in [n]: P_i \neq I} z_i \quad \forall\ P_1,\ \ldots,\ P_n \in \{I, X, Y, Z\}. \tag{43}$$

Let $z^* \in \mathbb{R}^n_{\geqslant 0}$ achieve the minimum in (43). 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

$$|c_{P_1 \ldots P_n}| = \sum_{i \in [n]: P_i \neq I} z^*_i. \tag{44}$$

From (43) we have $a \leqslant z^*_i$ for any $i = 1, \ldots, n$. Let $P_1, \ldots, P_n \in \{I, X, Y, Z\}$ satisfy (42), and let us assume that $w^*_{P_1 \ldots P_n} \neq 0$. We get from (44)

$$|c_{P_1 \ldots P_n}| = \sum_{i \in [n]: P_i \neq I} z^*_i \geqslant a\, |\{i \in [n] : P_i \neq I\}|, \tag{45}$$

which contradicts (42), and the claim follows. □

Appendix H. Supplementary details of estimated EM distance

Linear relaxation example over product states. To help aid intuition for the linear relaxation, consider the simple setting of estimating the quantum EM distance between two n-qubit product states $|\psi_1\rangle$ and $|\psi_2\rangle$ using only single qubit Pauli terms in the linear relaxation of (20) copied below:

$$\begin{aligned} &\text{maximize} \sum_{j=1}^{|W|} c_j w_j \\ &\text{subject to} \sum_{j:i\in\mathcal{I}_j} |w_j| \leqslant 1, \quad i = 1, \ldots, n \end{aligned} \tag{46}$$

where $c_j = \langle\psi_1|P_j|\psi_1\rangle - \langle\psi_2|P_j|\psi_2\rangle$ and P_j is one of the single qubit Pauli terms that we use in estimating the EM distance. Solving the above outputs an optimal Hamiltonian that takes the form

$$H_{\max} = \sum_{i=0}^{n_{\text{active}}} w^*_i H^*_i, \tag{47}$$

where w^*_i and H^*_i are the weights and active operators respectively.

In this setting, for each qubit i, the linear program above will set $w^*_j = \mathrm{sign}(c_j)$ for the c_j with largest magnitude in the set $\{j : i \in \mathcal{I}_j\}$ and $w^*_j = 0$ for all other elements in the set. Thus, the linear program will

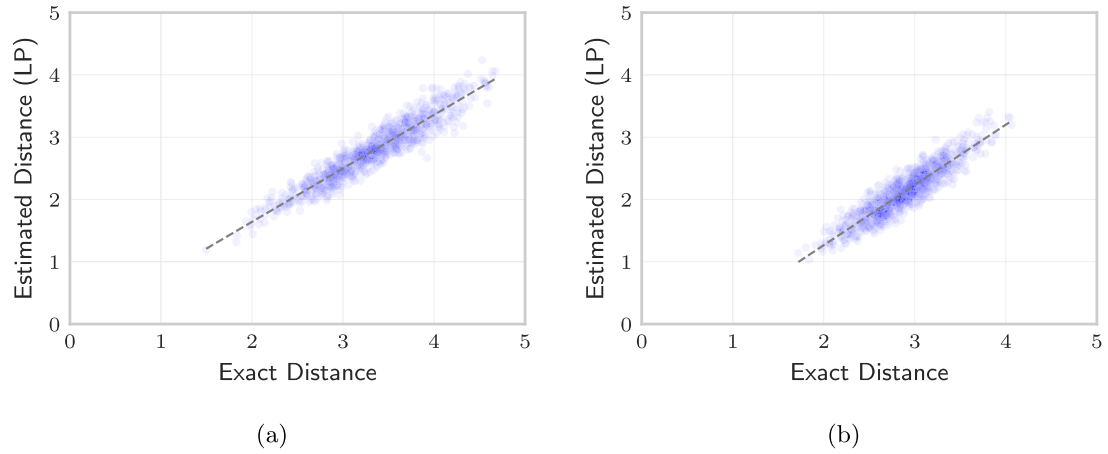

Figure 5. Randomly drawn values of the exact distance D_{EM} and estimated distance $D^{(2)}_{\text{EM}}$ (calculated from linear program over two-local Pauli operators) are highly correlated. Individual points are drawn from (a) random product states and (b) states drawn from randomly parameterizing three layers of the barren plateau mixing circuit (figure 12). All simulations are for a system of five qubits where exact calculations of the quantum EM distance can be performed efficiently on a classical computer. Each plot contains 1000 randomly drawn data points. A linear fit is plotted over the data as a dotted line.

select the Pauli terms that makes the largest contribution in the difference in expectations between the two states. For example, if $|\psi_1\rangle = |+\rangle|+\rangle$ and $|\psi_2\rangle = |-\rangle|-\rangle$, where $|+\rangle$ and $|-\rangle$ are the $+1$ and -1 eigenvectors of the X basis respectively, then the optimal Hamiltonian unsurprisingly equals $H_{\text{max}} = X \otimes I + I \otimes X$ since these are the single qubit Pauli terms that differ between the two states. Note, that $H_{\text{max}} = X \otimes I + I \otimes X$ also has a Lipschitz constant of one since any single qubit change can only change the value of the Hamiltonian by at most one.

As an aside, adding higher order Pauli terms in the above linear program makes no difference in the outcome. Note, that for any higher order Pauli term, the magnitude of the difference between expectations of the two states $|\psi_1\rangle$ and $|\psi_2\rangle$ will only be captured by differences in their individual qubits. Since the difference in expectation for this higher order Pauli term cannot exceed that of the single qubit Pauli terms within it, these terms will never appear in the optimal Hamiltonian calculated above (see appendix G for further details).

Correlation of estimated and actual quantum EM distance. Exact calculations of the quantum EM distance would require computational resources that grow exponentially with the number of qubits. However, as discussed in the main text, efficient estimates that lower bound the quantum EM distance can be obtained by formulating a new metric $D^{(k)}_{\text{EM}}$ optimizing over Hamiltonians of local operators (also see prior section for further motivation for this formalism). This estimated distance can be efficiently calculated as a linear program that requires computational resources that grow polynomially with the number of qubits.

Figure 5 shows that the exact distance D_{EM} and estimated distance $D^{(2)}_{\text{EM}}$ are well correlated for a system of five qubits where calculation of exact distances is possible on a classical computer. In figure 5, exact and estimated distances are compared for random product states (figure 5(a)) and states drawn randomly from a depth 3 circuit (figure 5(b)). Though correlations are strong for random product states (figure 5(a)), the estimated distance is not as strong of an approximator for the depth 3 circuit (figure 5(b)) where the slope of the correlation is slightly below one. This situation is one where many of the qubits have interacted with each other and we do not expect learning with the estimated distance to perform well at all times since the approximation is clearly not optimal (i.e., higher order Pauli operators may be needed to approximate the distance better). Nevertheless, the results here lend further support to the proof in the prior section showing that the optimal Hamiltonian in the quantum EM distance is biased towards local operators.

Appendix I. Gradients of qWGAN

For the generic version of our generator (equation (22)), optimization is performed over probability parameters p_i and gate parameters in each unitary U_i. The generator optimizes the parameters θ to minimize $\text{Tr}\,[G(\theta)\,H]$, where H is the Hamiltonian provided by the discriminator. The following proposition 7 proves that the gradient of $\text{Tr}\,[G(\theta)\,H]$ 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 7. *For any target quantum state σ, any parametric family of quantum states $\rho(t), 0 \leqslant t \leqslant T$ that is differentiable in $t = 0$ and any $k = 1, \ldots, n$,*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} D_{\text{EM}}(\rho(t), \sigma)\Big|_{t=0} &= \max\left(\text{Tr}\left[\rho'(0)\, H\right] : H \in \mathcal{O}_n,\ \|H\|_L \leqslant 1,\ \text{Tr}\left[(\rho(0) - \sigma)\, H\right] = D_{\text{EM}}(\rho(0), \sigma)\right),\\ \frac{\mathrm{d}}{\mathrm{d}t} D^{(k)}_{\text{EM}}(\rho(t), \sigma)\Big|_{t=0} &= \max\left(\text{Tr}\left[\rho'(0)\, H\right] : H \in \mathcal{O}^{(k)}_n,\ \|H\|_{\tilde{L}} \leqslant 1,\ \text{Tr}\left[(\rho(0) - \sigma)\, H\right] = D^{(k)}_{\text{EM}}(\rho(0), \sigma)\right).\end{aligned} \tag{48}$$

If $\rho(t)$ admits a differentiable extension to negative values of t, (48) provides the right derivative of $D_{\text{EM}}(\rho(t) - \sigma)$, which can be different from the left derivative if the max *in (48) is nontrivial.*

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

$$\frac{\partial D_{\text{EM}}}{\partial p_i} = \text{Tr}(U_i \rho_0 U_i^\dagger H_{\max}), \tag{49}$$

where $H_{\max}$ is the optimal Hamiltonian outputted by the discriminator (equation (21)). Note, that the above is simply the average measured value of $H_{\max}$ for the quantum state $U_i \rho_0 U_i^\dagger$. For gate parameters, we can use standard techniques [54] for evaluating gradients with respect to gate parameters.

I.1. Proof of proposition 7

We prove the claim for the exact quantum EM distance. The proof for the approximated quantum EM distance is completely analogous.

On the one hand, we have for any H as in (48)

$$\liminf_{t\to 0^+} \frac{\|\rho(t) - \sigma\|_{\text{EM}} - \|\rho(0) - \sigma\|_{\text{EM}}}{t} \geqslant \liminf_{t\to 0^+} \text{Tr}\left[\frac{\rho(t) - \rho(0)}{t} H\right] = \text{Tr}\left[\rho'(0)\, H\right]. \tag{50}$$

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

$$\limsup_{t\to 0^+} \frac{\|\rho(t) - \sigma\|_{\text{EM}} - \|\rho(0) - \sigma\|_{\text{EM}}}{t} \leqslant \limsup_{t\to 0^+} \text{Tr}\left[\frac{\rho(t) - \rho(0)}{t} H(t)\right]. \tag{51}$$

Let $t_k \downarrow 0$ be a sequence that achieves the lim sup in the right-hand side of (51) and such that

$$\lim_{k\to\infty} H(t_k) = H_0 \in \mathcal{O}_n. \tag{52}$$

We have

$$\begin{aligned}\|H_0\|_L &= \lim_{k\to\infty} \|H(t_k)\|_L \leqslant 1,\\ \text{Tr}\left[(\rho(0) - \sigma)\, H_0\right] &= \lim_{k\to\infty} \text{Tr}\left[(\rho(t_k) - \sigma)\, H(t_k)\right] = \lim_{k\to\infty} \|\rho(t_k) - \sigma\|_{\text{EM}} = \|\rho(0) - \sigma\|_{\text{EM}},\end{aligned} \tag{53}$$

and

$$\limsup_{t\to 0^+} \frac{\|\rho(t) - \sigma\|_{\text{EM}} - \|\rho(0) - \sigma\|_{\text{EM}}}{t} \leqslant \lim_{k\to\infty} \text{Tr}\left[\frac{\rho(t_k) - \rho(0)}{t_k} H(t_k)\right] = \text{Tr}\left[\rho'(0)\, H_0\right], \tag{54}$$

and the claim follows.

Appendix J. Additional simulations and figures

J.1. Learning the GHZ state

The analysis in section 4 showed that the qWGAN is especially effective at learning the GHZ state. In addition to the results shown in section 4, figure 6 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.

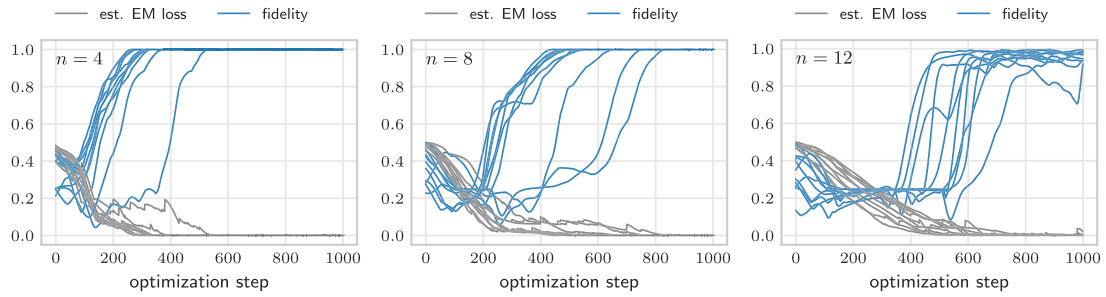

Figure 6. The qWGAN consistently generates the GHZ state in simulations with circuits of 4, 8, and 12 qubits. Estimated EM loss (quantum EM distance estimated by active operators) is also plotted in above chart, normalized by dividing by the number of qubits. Plots contain one line for each of 10 simulations for each circuit size.

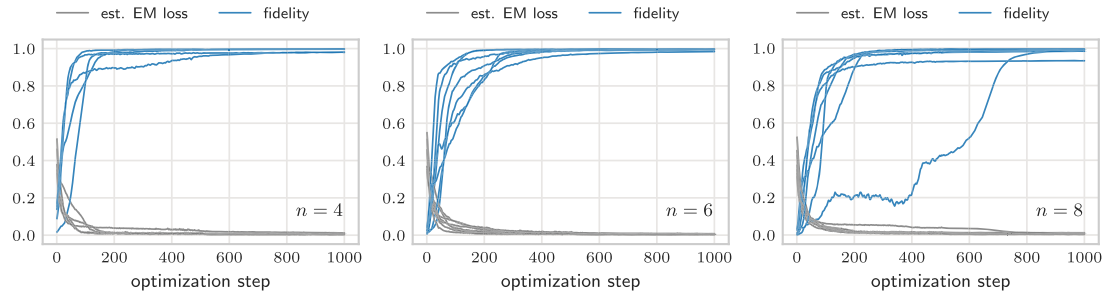

Figure 7. The student circuit is able to approximate well the state generated by the teacher circuit. Here, the target is constructed by randomly setting the parameters of a depth 2 mixing circuit (teacher circuit). The qWGAN, equipped with a generator circuit of depth 4, successfully learns the target state generated by the teacher circuit. For each plot, 5 simulations are performed.

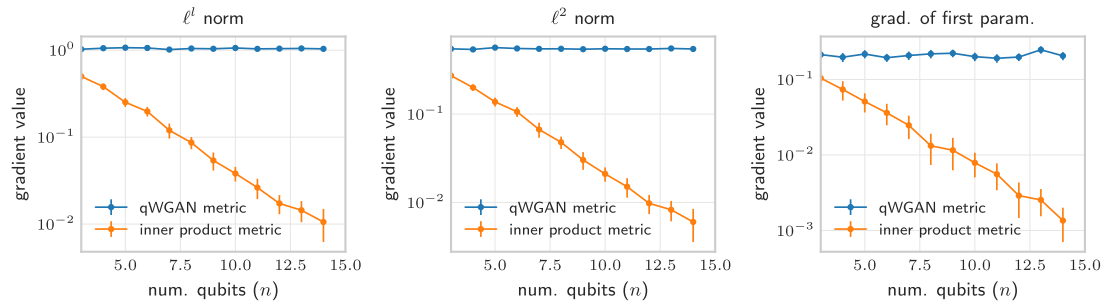

Figure 8. Comparison of gradients between the quantum EM loss metric and a conventional loss metric that is a function of the inner product. Here the average L_1 norm (left), L_2 norm (center), and the absolute value of the gradient of the first parameter (right) are shown. Decaying gradients observed 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.

J.2. Teacher–student learning

Supplementary to the results in section 4, we include figure 7 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.

J.3. Gradients of qWGAN vs. conventional GANs

Supplementary to figure 3(b), we include further details of the gradients of the quantum EM loss metric and its comparison to a inner product loss metric in figure 8. 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

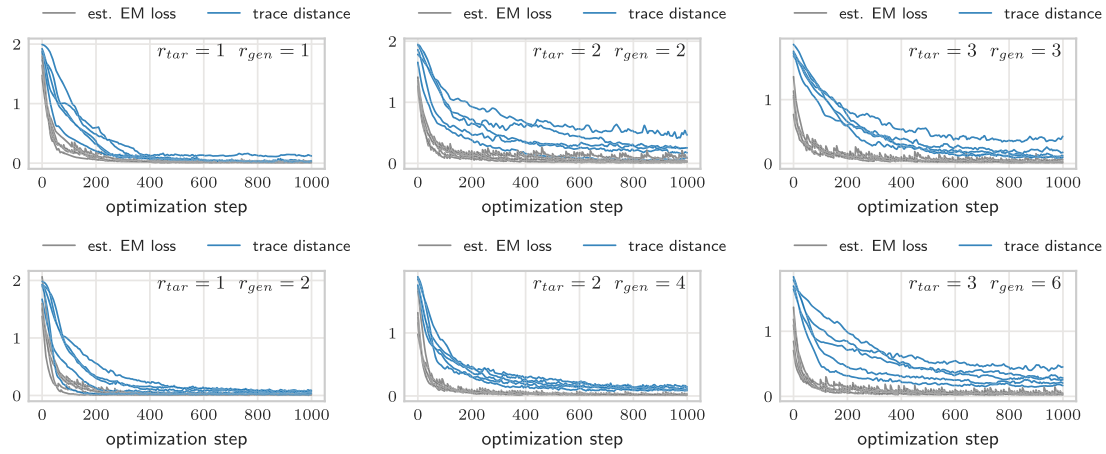

Figure 9. The qWGAN is also able to generate mixed states that approximate well their given target (also a mixed state) in both trace distance and quantum EM distance. Here, the generator circuit takes the form of a butterfly circuit of 4 qubits (see appendix K), and the target is constructed by randomly 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, five simulations are performed.

motivated by prior work in classical machine learning and photonics where similar parameterizations of unitary transformations produced interesting results [140–144]. 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,

$$\rho_{\text{tar}} = \frac{1}{r_{\text{tar}}} \sum_{i=1}^{r_{\text{tar}}} U_b(\theta_{\text{ran}}^{(i)}) \rho_0 U_b(\theta_{\text{ran}}^{(i)})^\dagger \tag{55}$$

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 9 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 (22) may not be well suited to optimization over mixed states. For example, it is often the case that different circuits in the generator optimize 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 QAOA, both given the acronym QAOA, are promising candidates for achieving quantum speedups in classical optimization problems [56–58]. Recent work has shown that QAOA is computationally universal [102] and potentially an effective algorithm in a wide range of quantum machine learning settings [11, 103, 145–148].

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:

$$C = B \sum_{i=1}^{N} \sigma_Z^{(i)} \sigma_Z^{(i+1)}, \tag{56}$$

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 $|10\rangle^{\otimes \frac{1}{2}n}$. Note that this setting is different from more traditional QAOA settings since here we do not aim to find the ground state but instead are given the ground state and aim to construct that state from a parameterized circuit.

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 $2L$

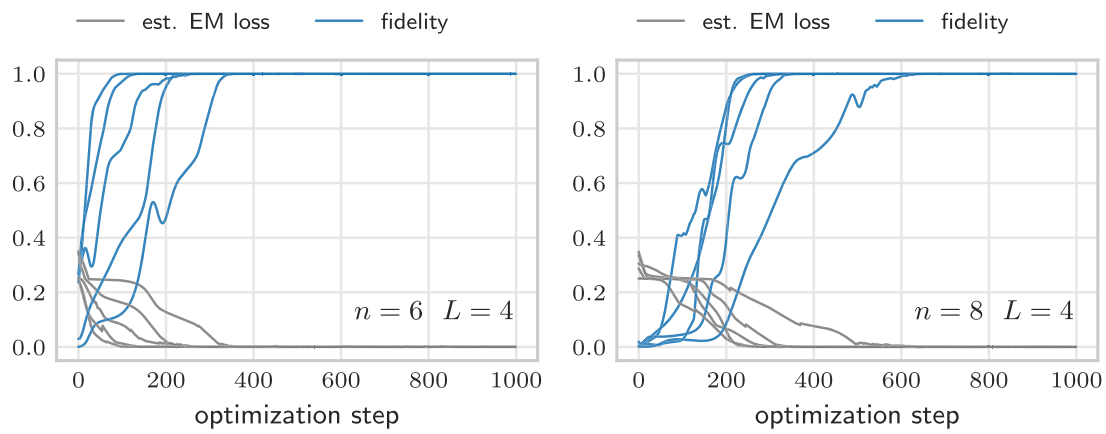

Figure 10. The qWGAN is effective at learning the ground state of a translationally invariant Ising Hamiltonian. Here, the generator is a QAOA circuit (see appendix K) of depth $L = 4$. Estimated W_1 loss (quantum EM distance estimated by active operators) is also plotted in above chart, normalized by dividing by the number of qubits. For each plot, 5 simulations are performed.

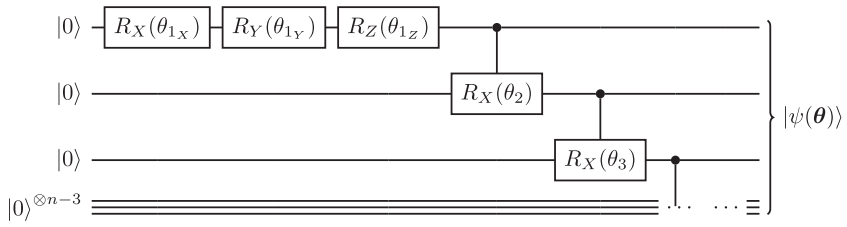

Figure 11. Circuit for generator in GHZ simulations (section 4).

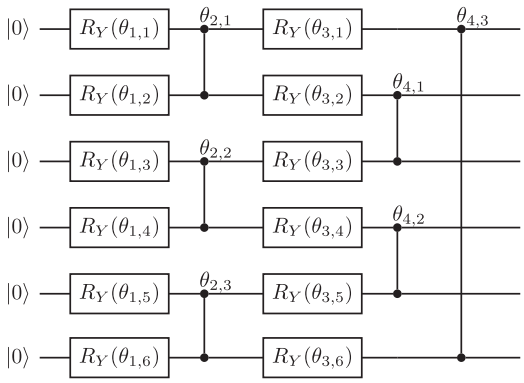

Figure 12. Single layer of mixing circuit used to perform learning and generate targets in section 4. This circuit consists of alternating layers of parameterized Pauli Y rotations and parameterized Pauli Z–Z rotations. The circuit above may be repeated to construct deeper circuits for simulations.

trainable parameters α_l and β_l (see appendix K for details of circuit).

$$H_{\text{mix}} = \sum_{i=1}^{N} \sigma_X^{(i)} \qquad H_{\text{C}} = \sum_{i=1}^{N} \sigma_Z^{(i)} \sigma_Z^{(i+1)} \tag{57}$$

Figure 10 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.

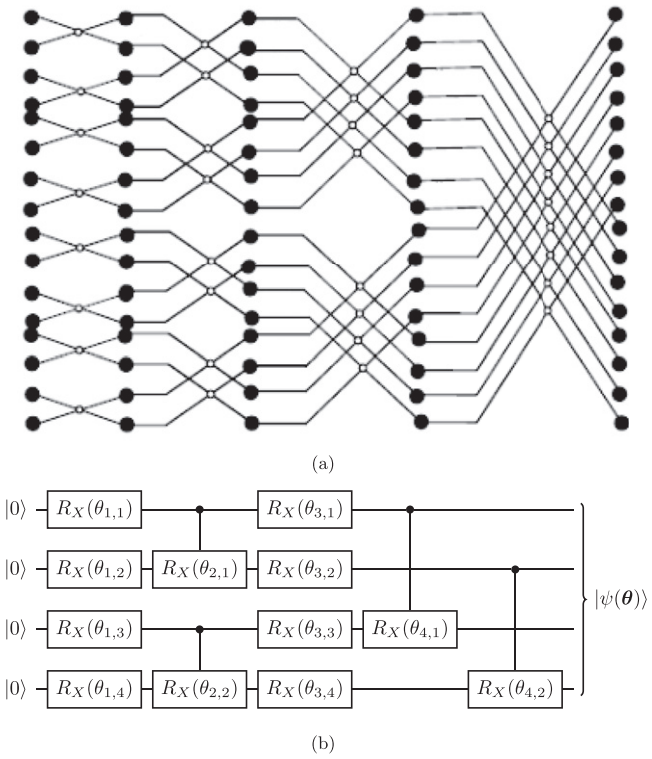

Figure 13. (a) Butterfly pattern of interactions, here shown for a system of 16 qubits. (b) Circuit for generator in butterfly circuit simulations (subsection J.4) here shown for 4 qubits.

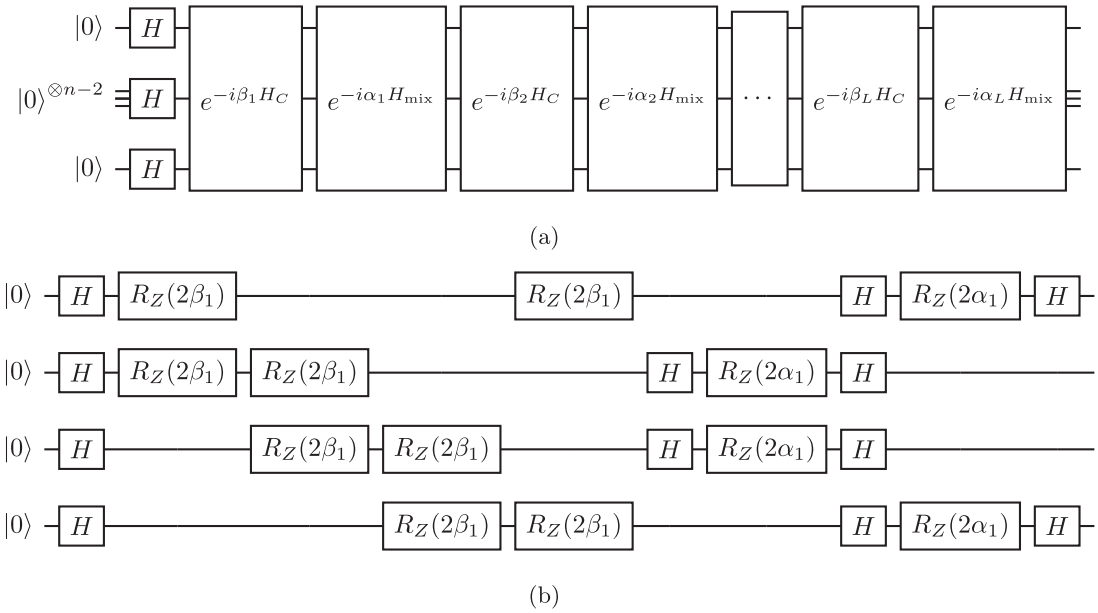

Figure 14. (a) General form for QAOA circuit. (b) QAOA circuit for 4 qubits and $L = 1$.

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 (section 4): circuit is shown in figure 11. This circuit differs from that used in the toy model (figure 1(a)) 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 (section 4): circuit is shown in figure 12. This circuit is commonly used in prior literature to show the existence of barren plateaus in the loss landscape [21, 23]. This circuit contains alternating layers of parameterized Pauli Y rotations and pairwise Pauli Z–Z rotations.
- 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 13(a)). The form of the circuit for 4 qubits shown in figure 13(b).
- QAOA circuit (subsection J.5): general form of circuit is shown in figure 14(a) 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 [149]. The form of the circuit for 4 qubits and a single QAOA layer ($L = 1$) is shown in figure 14(b).

Appendix L. Computational details

All code used for this paper is available here: https://github.com/bkiani/Quantum-EM-distance-and-qWGAN.

Quantum circuit simulations were performed using PennyLane [149] with a backend of Tensorflow [150] or Pytorch [151]. Unless specified otherwise, the Adam optimizer is used for performing gradient-based updates on a generator [59]. 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 $P = 0.8$ is used (see section 3.1). Discriminators are initialized with the set of two-local Pauli operators.

ORCID iDs

Bobak Toussi Kiani https://orcid.org/0000-0003-1477-0308
Giacomo De Palma https://orcid.org/0000-0002-5064-8695
Milad Marvian https://orcid.org/0000-0002-3049-6516
Zi-Wen Liu https://orcid.org/0000-0002-3402-9763

References

[1] Benedetti M, Grant E, Wossnig L and Severini S 2019 Adversarial quantum circuit learning for pure state approximation *New J. Phys.* **21** 043023
[2] Dallaire-Demers P-L and Killoran N 2018 Quantum generative adversarial networks *Phys. Rev.* A **98** 012324
[3] Torlai G and Melko R G 2020 Machine-learning quantum states in the NISQ era *Annu. Rev. Condens. Matter Phys.* **11** 325–44
[4] Gao J *et al* 2018 Experimental machine learning of quantum states *Phys. Rev. Lett.* **120** 240501
[5] Aaronson S 2007 The learnability of quantum states *Proc. R. Soc.* A **463** 3089–114
[6] Rocchetto A *et al* 2019 Experimental learning of quantum states *Sci. Adv.* **5** eaau1946
[7] Lloyd S and Weedbrook C 2018 Quantum generative adversarial learning *Phys. Rev. Lett.* **121** 040502
[8] Carrasquilla J, Torlai G, Melko R G and Aolita L 2019 Reconstructing quantum states with generative models *Nat. Mach. Intell.* **1** 155–61
[9] Chakrabarti S, Yiming H, Li T, Feizi S and Wu X 2019 Quantum Wasserstein generative adversarial networks *Advances in Neural Information Processing Systems* pp 6781–92
[10] Beer K *et al* 2020 Training deep quantum neural networks *Nat. Commun.* **11** 1–6
[11] Kiani B T, Lloyd S and Maity R 2020 Learning unitaries by gradient descent (arXiv:2001.11897)
[12] Mitarai K, Negoro M, Kitagawa M and Fujii K 2018 Quantum circuit learning *Phys. Rev.* A **98** 032309
[13] Bisio A, Chiribella G, D'Ariano G M, Facchini S and Perinotti P 2010 Optimal quantum learning of a unitary transformation *Phys. Rev.* A **81** 032324
[14] Quintino M T, Dong Q, Shimbo A, Soeda A and Murao M 2019 Reversing unknown quantum transformations: universal quantum circuit for inverting general unitary operations *Phys. Rev. Lett.* **123** 210502
[15] Lloyd S *et al* 2020 Quantum polar decomposition algorithm (arXiv:2006.00841)
[16] Carolan J *et al* 2020 Variational quantum unsampling on a quantum photonic processor *Nat. Phys.* **16** 322–7
[17] Sharma K, Khatri S, Cerezo M and Coles P J 2020 Noise resilience of variational quantum compiling *New J. Phys.* **22** 043006
[18] Benedetti M *et al* 2019 A generative modeling approach for benchmarking and training shallow quantum circuits *npj Quantum Inf.* **5** 1–9
[19] Liu J-G and Wang L 2018 Differentiable learning of quantum circuit born machines *Phys. Rev.* A **98** 062324
[20] Coyle B, Mills D, Danos V and Kashefi E 2020 The born supremacy: quantum advantage and training of an Ising born machine *npj Quantum Inf.* **6** 1–11

[21] McClean J R, Boixo S, Smelyanskiy V N, Babbush R and Neven H 2018 Barren plateaus in quantum neural network training landscapes *Nat. Commun.* **9** 1–6
[22] Wang S *et al* 2020 Noise-induced barren plateaus in variational quantum algorithms (arXiv:2007.14384)
[23] Cerezo M, Sone A, Volkoff T, Cincio L and Coles P J 2020 Cost-function-dependent barren plateaus in shallow quantum neural networks (arXiv:2001.00550)
[24] Pechen A N and Tannor D J 2011 Are there traps in quantum control landscapes? *Phys. Rev. Lett.* **106** 120402
[25] Moore K W and Rabitz H 2012 Exploring constrained quantum control landscapes *J. Chem. Phys.* **137** 134113
[26] Cerezo M *et al* 2020 Variational quantum algorithms (arXiv:2012.09265)
[27] De Palma G, Marvian M, Trevisan D and Lloyd S 2021 The quantum Wasserstein distance of order 1 *IEEE Trans. Inf. Theory* **67** 6627–43
[28] Arjovsky M, Chintala S and Bottou L 2017 Wasserstein GAN (arXiv:1701.07875)
[29] Chen L *et al* 2018 Adversarial text generation via feature-mover's distance *Advances in Neural Information Processing Systems* pp 4666–77
[30] Rubner Y, Tomasi C and Guibas L J 1998 A metric for distributions with applications to image databases *6th Int. Conf. Computer Vision (IEEE Cat. No. 98CH36271)* (Piscataway, NJ: IEEE) pp 59–66
[31] Villani C 2008 *Optimal Transport: Old and New* vol 338 (Berlin: Springer)
[32] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C 2017 Improved training of Wasserstein GANS *Advances in Neural Information Processing Systems* pp 5767–77
[33] Goodfellow I *et al* 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* pp 2672–80
[34] Hu L *et al* 2019 Quantum generative adversarial learning in a superconducting quantum circuit *Sci. Adv.* **5** eaav2761
[35] Campos E, Nasrallah A and Biamonte J 2021 Abrupt transitions in variational quantum circuit training *Phys. Rev.* A **103** 032607
[36] Skolik A, McClean J R, Mohseni M, van der Smagt P and Leib M 2020 Layerwise learning for quantum neural networks (arXiv:2006.14904)
[37] Brandao F G and Svore K M 2017 Quantum speed-ups for solving semidefinite programs *2017 IEEE 58th Annual Symp. Foundations of Computer Science (FOCS)* (Piscataway, NJ: IEEE) pp 415–26
[38] Van Apeldoorn J, Gilyén A, Gribling S and de Wolf R 2020 Quantum SDP-solvers: better upper and lower bounds *Quantum* **4** 230
[39] Bertsimas D and Tsitsiklis J N 1997 *Introduction to Linear Optimization* vol 6 (Belmont, MA: Athena Scientific)
[40] Huang H-Y, Kueng R and Preskill J 2020 Predicting many properties of a quantum system from very few measurements *Nat. Phys.* **16** 1050–7
[41] Huang H-Y, Kueng R and Preskill J 2021 Efficient estimation of Pauli observables by derandomization (arXiv:2103.07510)
[42] Krizhevsky A, Sutskever I and Hinton G E 2017 Imagenet classification with deep convolutional neural networks *Commun. ACM* **60** 84–90
[43] Gu J *et al* 2018 Recent advances in convolutional neural networks *Pattern Recognit.* **77** 354–77
[44] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[45] Vaswani A *et al* 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5998–6008
[46] Brown T B *et al* 2020 Language models are few-shot learners (arXiv:2005.14165)
[47] Devlin J, Chang M-W, Lee K and Toutanova K 2018 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)
[48] Benedetti M, Lloyd E, Sack S and Fiorentini M 2019 Parameterized quantum circuits as machine learning models *Quantum Sci. Technol.* **4** 043001
[49] Du Y, Hsieh M-H, Liu T and Tao D 2018 The expressive power of parameterized quantum circuits (arXiv:1810.11922)
[50] Sharma K, Cerezo M, Cincio L and Coles P J 2020 Trainability of dissipative perceptron-based quantum neural networks (arXiv:2005.12458)
[51] Schuld M, Sinayskiy I and Petruccione F 2014 The quest for a quantum neural network *Quantum Inf. Process.* **13** 2567–86
[52] Killoran N *et al* 2019 Continuous-variable quantum neural networks *Phys. Rev. Res.* **1** 033063
[53] Cong I, Choi S and Lukin M D 2019 Quantum convolutional neural networks *Nat. Phys.* **15** 1273–8
[54] Schuld M, Bergholm V, Gogolin C, Izaac J and Killoran N 2019 Evaluating analytic gradients on quantum hardware *Phys. Rev.* A **99** 032331
[55] Huembeli P and Dauphin A 2020 Characterizing the loss landscape of variational quantum circuits (arXiv:2008.02785)
[56] Fingerhuth M *et al* 2018 A quantum alternating operator ansatz with hard and soft constraints for lattice protein folding (arXiv:1810.13411)
[57] Hadfield S, Wang Z, O'Gorman B, Rieffel E, Venturelli D and Biswas R 2019 From the quantum approximate optimization algorithm to a quantum alternating operator ansatz *Algorithms* **12** 34
[58] Farhi E, Goldstone J and Gutmann S 2014 A quantum approximate optimization algorithm (arXiv:1411.4028)
[59] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[60] Ge X, Ding H, Rabitz H and Wu R-B 2020 Robust quantum control in games: an adversarial learning approach *Phys. Rev.* A **101** 052317
[61] Palittapongarnpim P, Wittek P, Zahedinejad E, Vedaie S and Sanders B C 2017 Learning in quantum control: high-dimensional global optimization for noisy quantum dynamics *Neurocomputing* **268** 116–26
[62] Schuld M, Bocharov A, Svore K M and Wiebe N 2020 Circuit-centric quantum classifiers *Phys. Rev.* A **101** 032308
[63] Romero J, Olson J P and Aspuru-Guzik A 2017 Quantum autoencoders for efficient compression of quantum data *Quantum Sci. Technol.* **2** 045001
[64] Jones T and Benjamin S C 2018 Quantum compilation and circuit optimisation via energy dissipation (arXiv:1811.03147)
[65] Nautrup H P, Delfosse N, Dunjko V, Briegel H J and Friis N 2019 Optimizing quantum error correction codes with reinforcement learning *Quantum* **3** 215
[66] Baireuther P, O'Brien T E, Tarasinski B and Beenakker C W J 2018 Machine-learning-assisted correction of correlated qubit errors in a topological code *Quantum* **2** 48
[67] Bausch J and Leditzky F 2020 Quantum codes from neural networks *New J. Phys.* **22** 023005
[68] Johnson P D, Romero J, Olson J, Cao Y and Aspuru-Guzik A 2017 Qvector: an algorithm for device-tailored quantum error correction (arXiv:1711.02249)
[69] Zhao C and Gao X-S 2021 Analyzing the barren plateau phenomenon in training quantum neural networks with the *zx*-calculus *Quantum* **5** 466

[70] Cerezo M *et al* 2021 Variational quantum algorithms *Nat. Rev. Phys.* **3** 625–44
[71] Larocca M, Calzetta E A and Wisniacki D A 2020 Navigating on quantum control solution subspaces (arXiv:2001.05941)
[72] Grant E, Wossnig L, Ostaszewski M and Benedetti M 2019 An initialization strategy for addressing barren plateaus in parameterized quantum circuits *Quantum* **3** 214
[73] Zhou L, Wang S-T, Choi S, Pichler H and Lukin M D 2020 Quantum approximate optimization algorithm: performance, mechanism, and implementation on near-term devices *Phys. Rev.* X **10** 021067
[74] Pesah A *et al* 2020 Absence of barren plateaus in quantum convolutional neural networks (arXiv:2011.02966)
[75] Bharti K and Haug T 2020 Quantum assisted simulator (arXiv:2011.06911)
[76] Stokes J, Izaac J, Killoran N and Carleo G 2020 Quantum natural gradient *Quantum* **4** 269
[77] Zhang H, Goodfellow I, Metaxas D and Odena A 2019 Self-attention generative adversarial networks *Int. Conf. Machine Learning (PMLR)* pp 7354–63
[78] Miyato T, Kataoka T, Koyama M and Yoshida Y 2018 Spectral normalization for generative adversarial networks (arXiv:1802.05957)
[79] Karras T, Aila T, Laine S and Lehtinen J 2017 Progressive growing of GANS for improved quality, stability, and variation (arXiv:1710.10196)
[80] Roth K, Lucchi A, Nowozin S and Hofmann T 2017 Stabilizing training of generative adversarial networks through regularization *Advances in Neural Information Processing Systems* pp 2018–28
[81] Petzka H, Fischer A and Lukovnicov D 2017 On the regularization of Wasserstein GANS (arXiv:1709.08894)
[82] Gao R, Chen X and Kleywegt A J 2017 Wasserstein distributional robustness and regularization in statistical learning (arXiv:1712.06050)
[83] Li Z, Meier M-A, Hauksson E, Zhan Z and Andrews J 2018 Machine learning seismic wave discrimination: application to earthquake early warning *Geophys. Res. Lett.* **45** 4773–9
[84] Xuan Q *et al* 2018 Multiview generative adversarial network and its application in pearl classification *IEEE Trans. Ind. Electron.* **66** 8244–52
[85] Yi Z, Zhang H, Tan P and Gong M 2017 Dualgan: unsupervised dual learning for image-to-image translation *Proc. IEEE Int. Conf. Computer Vision* pp 2849–57
[86] Elgammal A, Liu B, Elhoseiny M and Mazzone M 2017 Can: creative adversarial networks, generating art by learning about styles and deviating from style norms (arXiv:1706.07068)
[87] Wang Z, Wang J and Wang Y 2018 An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition *Neurocomputing* **310** 213–22
[88] Anand A, Romero J, Degroote M and Aspuru-Guzik A 2020 Experimental demonstration of a quantum generative adversarial network for continuous distributions (arXiv:2006.01976)
[89] Ahmed S, Muñoz C S, Nori F and Kockum A F 2020 Quantum state tomography with conditional generative adversarial networks (arXiv:2008.03240)
[90] Lu S, Duan L-M and Deng D-L 2020 Quantum adversarial machine learning *Phys. Rev. Res.* **2** 033212
[91] Zeng J, Wu Y, Liu J-G, Wang L and Hu J 2019 Learning and inference on generative adversarial quantum circuits *Phys. Rev.* A **99** 052306
[92] Romero J and Aspuru-Guzik A 2019 Variational quantum generators: generative adversarial quantum machine learning for continuous distributions (arXiv:1901.00848)
[93] Zoufal C, Lucchi A and Woerner S 2019 Quantum generative adversarial networks for learning and loading random distributions *npj Quantum Inf.* **5** 1–9
[94] Nakaji K and Yamamoto N 2020 Quantum semi-supervised generative adversarial network for enhanced data classification (arXiv:2010.13727)
[95] Herr D, Obert B and Rosenkranz M 2020 Anomaly detection with variational quantum generative adversarial networks (arXiv:2010.10492)
[96] Huang B, Symonds N O and von Lilienfeld O A 2020 Quantum machine learning in chemistry and materials *Handbook of Materials Modeling: Methods: Theory and Modeling* (New York: Springer) pp 1883–909
[97] Stamatopoulos N, Egger D J, Sun Y, Zoufal C, Iten R, Shen N and Woerner S 2020 Option pricing using quantum computers *Quantum* **4** 291
[98] Orús R, Mugel S and Lizaso E 2019 Quantum computing for finance: overview and prospects *Rev. Phys.* **4** 100028
[99] Kiani B T, Villanyi A and Lloyd S 2020 Quantum medical imaging algorithms (arXiv:2004.02036)
[100] Kiani B T *et al* 2020 Quantum advantage for differential equation analysis (arXiv:2010.15776)
[101] Yao X-W *et al* 2017 Quantum image processing and its application to edge detection: theory and experiment *Phys. Rev.* X **7** 031041
[102] Lloyd S 2018 Quantum approximate optimization is computationally universal (arXiv:1812.11075)
[103] Zhang Y, Zhang R and Potter A C 2020 QED driven QAOA for network-flow optimization (arXiv:2006.09418)
[104] Kandala A, Mezzacapo A, Temme K, Takita M, Brink M, Chow J M and Gambetta J M 2017 Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets *Nature* **549** 242–6
[105] Parrish R M, Hohenstein E G, McMahon P L and Martínez T J 2019 Quantum computation of electronic transitions using a variational quantum eigensolver *Phys. Rev. Lett.* **122** 230401
[106] Monge G 1781 *Mémoire sur la théorie des déblais et des remblais Meémoires de l'Acadeémie royale des sciences de Paris* vol **1781** pp 625–704
[107] Kantorovich L V 2006 On the translocation of masses *J. Math. Sci.* **133** 1381–2
[108] Ambrosio L, Gigli N and Savaré G 2008 *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Berlin: Springer)
[109] Peyré G and Cuturi M 2019 Computational optimal transport: with applications to data science *FNT Mach. Learn.* **11** 355–607
[110] Vershik A M 2013 Long history of the Monge–Kantorovich transportation problem *Math. Intell.* **35** 1–9
[111] Ornstein D S 1973 An application of ergodic theory to probability theory *Ann. Probab.* **1** 43–58
[112] Nielsen M A and Chuang I 2002 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[113] Carlen E A and Maas J 2014 An analog of the two-Wasserstein metric in non-commutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy *Commun. Math. Phys.* **331** 887–926
[114] Carlen E A and Maas J 2017 Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance *J. Funct. Anal.* **273** 1810–69

[115] Carlen E A and Maas J 2020 Non-commutative calculus, optimal transport and functional inequalities in dissipative quantum systems *J. Stat. Phys.* **178** 319–78
[116] Rouzé C and Datta N 2019 Concentration of quantum states from quantum functional and transportation cost inequalities *J. Math. Phys.* **60** 012202
[117] Datta N and Rouzé C 2020 Relating relative entropy, optimal transport and Fisher information: a quantum HWI inequality *Ann. Henri Poincare* **21** 2115–50
[118] Van Vu T and Hasegawa Y 2020 Geometrical bounds of the irreversibility in Markovian systems (arXiv:2005.02871)
[119] De Palma G and Huber S 2018 The conditional entropy power inequality for quantum additive noise channels *J. Math. Phys.* **59** 122201
[120] Gao L, Junge M and LaRacuente N 2020 Fisher information and logarithmic sobolev inequality for matrix-valued functions *Ann. Henri Poincare* **21** 3409–78
[121] Chen Y, Georgiou T T, Ning L and Tannenbaum A 2017 Matricial Wasserstein-1 distance *IEEE Control Syst. Lett.* **1** 14–9
[122] Ryu E K, Chen Y, Li W and Osher S 2018 Vector and matrix optimal mass transport: theory, algorithm, and applications *SIAM J. Sci. Comput.* **40** A3675–98
[123] Chen Y, Georgiou T T and Tannenbaum A 2018 Matrix optimal mass transport: a quantum mechanical approach *IEEE Trans. Autom. Control* **63** 2612–9
[124] Chen Y, Georgiou T T and Tannenbaum A 2018 Wasserstein geometry of quantum states and optimal transport of matrix-valued measures *Emerging Applications of Control and Systems Theory* (Berlin: Springer) pp 139–50
[125] Agredo J 2013 A Wasserstein-type distance to measure deviation from equilibrium of quantum Markov semigroups *Open Syst. Inf. Dyn.* **20** 1350009
[126] Agredo J 2016 On exponential convergence of generic quantum Markov semigroups in a Wasserstein-type distance *Int. J. Pure Appl. Math.* **107** 909–25
[127] Ikeda K 2020 Foundation of quantum optimal transport and applications *Quantum Inf. Process.* **19** 25
[128] Golse F, Mouhot C and Paul T 2016 On the mean field and classical limits of quantum mechanics *Commun. Math. Phys.* **343** 165–205
[129] Caglioti E, Golse F and Paul T 2021 Towards optimal transport for quantum densities (arXiv: 2101.03256)
[130] Golse F 2018 The quantum *N*-body problem in the mean-field and semiclassical regime *Phil. Trans. R. Soc.* A **376** 20170229
[131] Golse F and Paul T 2017 The Schrödinger equation in the mean-field and semiclassical regime *Arch. Ration. Mech. Anal.* **223** 57–94
[132] Golse F and Paul T 2018 Wave packets and the quadratic Monge–Kantorovich distance in quantum mechanics *C. R. Math.* **356** 177–97
[133] Caglioti E, Golse F and Paul T 2020 Quantum optimal transport is cheaper *J. Stat. Phys.* **181** 149–62
[134] De Palma G and Trevisan D 2021 Quantum optimal transport with quantum channels *Ann. Henri Poincare* **22** 3199–234
[135] Duvenhage R and Snyman M 2018 Balance between quantum Markov semigroups *Ann. Henri Poincare* **19** 1747–86
[136] Agredo J and Fagnola F 2017 On quantum versions of the classical Wasserstein distance *Stochastics* **89** 910–22
[137] Zyczkowski K and Slomczynski W 1998 The Monge distance between quantum states *J. Phys. A: Math. Gen.* **31** 9095
[138] Zyczkowski K and Slomczynski W 2001 The Monge metric on the sphere and geometry of quantum states *J. Phys. A: Math. Gen.* **34** 6689
[139] Bengtsson I and Życzkowski K 2017 *Geometry of Quantum States: An Introduction to Quantum Entanglement* (Cambridge: Cambridge University Press)
[140] Mathieu M and LeCun Y 2014 Fast approximation of rotations and Hessians matrices (arXiv:1404.7195)
[141] Jing L *et al* 2017 Tunable efficient unitary neural networks (EUNN) and their application to RNNS *Int. Conf. Machine Learning (PMLR)* pp 1733–41
[142] Dao T, Gu A, Eichhorn M, Rudra A and Ré C 2019 Learning fast algorithms for linear transforms using butterfly factorizations *Proc. Machine Learning Research* vol 97 p 1517
[143] Clements W R, Humphreys P C, Metcalf B J, Kolthammer W S and Walsmley I A 2016 Optimal design for universal multiport interferometers *Optica* **3** 1460–5
[144] Shen Y *et al* 2017 Deep learning with coherent nanophotonic circuits *Nat. Photon.* **11** 441
[145] Verdon G, Broughton M and Biamonte J 2017 A quantum algorithm to train neural networks using low-depth circuits (arXiv:1712.05304)
[146] Wang Z, Hadfield S, Jiang Z and Rieffel E G 2018 Quantum approximate optimization algorithm for maxcut: a fermionic view *Phys. Rev.* A **97** 022304
[147] Hodson M, Ruck B, Ong H, Garvin D and Dulman S 2019 Portfolio rebalancing experiments using the quantum alternating operator ansatz (arXiv:1911.05296)
[148] Chancellor N 2019 Domain wall encoding of discrete variables for quantum annealing and QAOA *Quantum Sci. Technol.* **4** 045004
[149] Bergholm V *et al* 2018 Pennylane: automatic differentiation of hybrid quantum–classical computations (arXiv:1811.04968)
[150] Abadi M *et al* 2015 TensorFlow: large-scale machine learning on heterogeneous systems (https://tensorflow.org/) Software available from (tensorflow.org)
[151] Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* pp 8026–37