Quantum Information & Computation
ISSN 1533-7146 Volume 25 (2025) pp. 156–174
©Cerebration Science Publishing

Article

Polynomial Complexity of Quantum Sample Tomography

Kun Tang and Jun Lai *

School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China; kun.tang@zju.edu.cn
* Correspondence author: laijun6@zju.edu.cn

Received date: 9 February 2025; Accepted date: 11 March 2025; Published online: 22 March 2025

Abstract: Efficient quantum tomography is crucial for advancing quantum computing technologies. Traditional quantum state tomography requires an exponential number of measurements for complete reconstruction. Therefore, developing methods that reduce measurement complexity to polynomial scale is essential for practical applications. In this paper, we show that quantum sample tomography can be accomplished with polynomial scale measurements while maintaining accuracy with a high probability. We present a novel approach using conditional recurrent neural networks (RNNs) with solid theoretical foundations from Rademacher complexity and random projection theory. The effectiveness of our method is validated through several quantum models.

Keywords: quantum sample tomography; Rademacher complexity; random projection; neural network

1. Introduction

Quantum computing represents a revolutionary computational paradigm by offering substantial advantages in various applications, including quantum system simulation [1], linear equations [2], and combinatorial optimization [3]. Although fully functional quantum computers remain unrealized, their potential impact on computational capabilities is profound. One of the big challenges in quantum computing is figuring out how to reconstruct quantum states. Specifically, a system with n qubits requires approximately 4^n measurements for complete state reconstruction, resulting in exponential computational costs. For example, reconstructing a 10-qubit system requires 8,192 measurements across 59,049 distinct quantum circuits, requiring roughly 130 hours on IBM quantum processors [4]. Consequently, developing efficient methods for quantum state reconstruction is crucial for practical quantum computing applications.

Several approaches have been developed to address the challenges of quantum state reconstruction. These include maximum likelihood estimation [5], projected gradient descent [6], and compressed sensing [7]. For instance, compressed sensing reduces the computational complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d \log d)$, where $d = 2^n$ represents the density matrix dimension. Some works also address quantum state tomography for specific cases, such as matrix product density operators [8] and low rank density matrix [9]. While these methods have demonstrated promising results, they primarily focus on full density matrix reconstruction and fail to achieve exponential acceleration. Furthermore, their reliance on iterative procedures complicates both convergence analysis and complexity estimation.

Recently, neural network-based approaches have emerged as promising tools to quantum state reconstruction challenges. Initially started from the application of restricted Boltzmann machines (RBM) [10], this field has expanded to incorporate deep learning methods [11,12], adaptive measurement base selection techniques [13], and quantum machine learning methods [14]. These approaches compress quantum states into statistical learning models, enabling information extraction without exhaustive measurements. A key advantage of these methods is their ability to infer measurement outcomes from partial data, eliminating the need for complete state reconstruction. For example, when computing the trace of the product between a quantum state density matrix ρ of n-qubit and a given matrix M, full density matrix reconstruction can be avoided if M can be decomposed as:

$$M = \sum_{i=1}^{k} \alpha_i M_i,$$

where k is significantly smaller than $O(4^n)$ and $tr(\rho M_i)$ is computable in polynomial time.

DOI: 10.2478/qic-2025-0008

Despite these algorithmic advances, there are very few theoretical justifications. An important theoretical foundation for neural network applications in quantum state reconstruction was established by Aaronson [15], who introduced the concept of quantum shadow tomography. While his work demonstrated that two-outcome measurements could be predicted using polynomial scale sampling, it left open the question of whether it is possible to predict distributional measurement outcomes, similar to classical shadow tomography [16,17], with polynomial complexity. In this paper, we extend this theoretical framework by employing local Rademacher complexity analysis to prove the effectiveness of quantum sample tomography for measurements with distributional outcomes. Relying on the theorem from Bartlett et al. [18], we demonstrate that the required number of measurements scales polynomially with qubit number, which is shown in Theorem 1. By incorporating random projection techniques, we show that this relationship holds even for high-dimensional quantum systems. To validate our findings, we conduct numerical experiments using RNNs. We configure the sample size and the number of RNN parameters to scale polynomially with qubit number and introduce a penalty term in the loss function to satisfy the conditions of our theorem. Our experiments on the ground state of the transverse field Ising model (TFIM) show agreement between theoretical predictions and numerical results. Furthermore, we compare our RNN approach against other models and show its effectiveness in various quantum states including cat states, random states, and W states.

The remainder of this paper is structured as follows. In Section 2, we introduce the fundamentals of quantum sample tomography and provide the main result. Sections 3 and 4 address the theoretical challenges using our new framework and give a rigorous proof of the main theorem. In Sections 5 and 6, we validate the theoretical findings through several quantum models, with particular emphasis on the performance of the proposed neural network. Section 7 concludes the paper.

2. Preliminary

We first review the fundamental concepts in quantum computation and give some necessary notations.

A quantum system is mathematically represented by a density matrix ρ, which is characterized by three essential properties: Hermiticity ($\rho = \rho^\dagger$), positive definiteness ($\rho \succeq 0$), and unit trace ($\mathrm{tr}\rho = 1$). The matrix elements may take complex values, reflecting the quantum mechanical nature of the system. Under physical operations implemented through quantum circuits, the system undergoes unitary transformations: when a unitary matrix U is applied, the density matrix transforms as $\rho \to U\rho U^\dagger$. Modern quantum computers, despite their varied physical implementations, fundamentally operate by applying sequences of unitary transformations. According to the Solovay-Kitaev theorem [19], arbitrary unitary operations can be efficiently approximated to any desired precision using a finite universal set of elementary quantum gates.

A fundamental challenge in quantum state characterization is the reconstruction of an unknown quantum state ρ. The standard approach to this problem, known as quantum state tomography, relies on measurements using Pauli operators. The Pauli matrices are defined as follows:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The Pauli matrices, together with the identity matrix I, constitute a complete basis for two-dimensional Hermitian matrices. In quantum mechanical measurements, the expectation value of a Pauli operator M with respect to a quantum state ρ is given by $\mathrm{tr}(\rho M)$.

The measurement principles extend naturally to multi-qubit systems, where measurements can be performed on individual qubits. For n-qubit systems, the density matrix can be expressed in terms of Pauli operator measurements through the following decomposition:

$$\rho = \sum_{\{v_1,\ldots,v_n\}\in 4^{\otimes n}} \frac{\mathrm{tr}(\rho\sigma_1^{v_1}\otimes\sigma_2^{v_2}\otimes\cdots\sigma_n^{v_n})\sigma_1^{v_1}\otimes\sigma_2^{v_2}\otimes\cdots\sigma_n^{v_n}}{2^n},$$

where $\sigma_i^{v_i}$ denotes the Pauli operator on the i-th qubit with $v_i \in \{1,2,3,4\}$, corresponding to $\{\sigma_i^1 = I, \sigma_i^2 = X, \sigma_i^3 = Y, \sigma_i^4 = Z\}$. The complete reconstruction of the quantum state requires measuring all terms in this decomposition, requiring at least $O(3^n)$ distinct measurements [20]. This exponential scaling with respect to qubit number presents a fundamental challenge for quantum state tomography of large-scale quantum systems.

To address the exponential scaling challenge, Aaronson [15] introduced the concept of shadow tomography. The problem can be formally stated as follows:

Problem 1. *Given an unknown $D = 2^n$ dimensional quantum mixed state ρ and a distribution $\mathcal{D}$ over two-outcome measurements, we sample m measurements $\{E_1, \ldots, E_m\}$ independently from $\mathcal{D}$ and obtain their measurement results $b_i = \mathrm{tr}(\rho E_i)$ for $i = 1, \ldots, m$. For any chosen failure probability $\delta > 0$, accuracy $\gamma > 0$, and error tolerance $\varepsilon > 0$, what is the minimum number of measurements m (depending on D, ε, and δ), such that any hypothesis state σ minimizing the quadratic functional $\sum_{i=1}^{m} (\mathrm{tr}(\sigma E_i) - b_i)^2$ satisfies*

$$\mathrm{Pr}_{E\in\mathcal{D}}(|\mathrm{tr}(\rho E) - \mathrm{tr}(\sigma E)| > \gamma) \leq \varepsilon$$

with probability at least $1 - \delta$?

This formulation essentially asks whether a statistical learning model exists that can extract exponentially large amounts of information from a relatively small dataset. Aaronson provided an affirmative answer by proving that only

$$m_{min} = \mathcal{O}\left(\frac{1}{\gamma^2\varepsilon^2}\left(\frac{n}{\gamma^2\varepsilon^2}\ln^2\left(\frac{1}{\gamma\varepsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

measurements are sufficient, given that $\gamma\varepsilon \geq 7\eta$, with $\eta > 0$ being the upper bound of $|\mathrm{tr}(\sigma E_i) - \mathrm{tr}(\rho E_i)|$ for all $i = 1, \ldots, m$. This result demonstrates that approximating an exponential number of two-outcome measurements requires only a number of quantum state measurements that scale linearly with n, offering a considerable improvement over the exponential complexity of traditional approaches.

While Aaronson's result focuses on predicting two-outcome measurements, our work addresses quantum sample tomography, building on the experimental investigations by Smith et al. [4]. In the subsequent analysis, let $\mathbb{E}$ denote the expectation. The problem can be formally defined as follows:

Problem 2. *Given an unknown $D = 2^n$ dimensional quantum mixed state ρ and a distribution $\mathcal{D}$ over two-outcome measurements, we sample m unitaries $\{U_1, \ldots, U_m\}$ independently from $\mathcal{D}$ and obtain their measurement results $b_i = \mathrm{diag}(U_i^\dagger \rho U_i)$ for $i = 1, \ldots, m$. For any chosen failure probability $\delta > 0$ and error tolerance $\varepsilon > 0$, what is the minimum number of measurements m (depending on D, ε, and δ) such that any hypothesis state σ minimizing the quadratic functional $\sum_{i=1}^{m} \|\mathrm{diag}(U_i^\dagger \sigma U_i) - b_i\|^2$, where $\|\cdot\|$ is given in the sense of l_2 norm, satisfies*

$$\mathbb{E}_{U\in\mathcal{D}}(\|\mathrm{diag}(U^\dagger \sigma U) - \mathrm{diag}(U^\dagger \rho U)\|) \leq \varepsilon$$

with probability at least $1 - \delta$?

While structurally similar to Problem 1, our formulation differs in three key aspects. First, instead of two-outcome measurements, we consider unitary transformations as independent variables. Second, our output consists of complete probability distributions, which can be represented as vectors in practical implementations, rather than binary outcomes. Third, we employ expectation rather than probability, which is more general. In fact, using Markov's inequality, if we choose $\varepsilon = \varepsilon'\gamma'$, where $\varepsilon', \gamma' > 0$, then it holds

$$\mathrm{Pr}_{U\in\mathcal{D}}(\|\mathrm{diag}(U^\dagger \sigma U) - \mathrm{diag}(U^\dagger \rho U)\| > \gamma') \leq \varepsilon',$$

which is consistent with the conclusion of Problem 1. Thus, our formulation represents a natural generalization of Aaronson's original shadow tomography problem.

In the subsequent analysis, we let the loss function $\ell(x, y) = \|x - y\|_2^2$ unless otherwise specified. We denote $poly(n)$ a polynomial function of variable n. We provide an affirmative answer to Problem 2 under the assumption that ρ is sparse, which is stated in the following theorem:

Theorem 1. *Let ρ be an n-qubit quantum state and $\mathcal{D}$ be a distribution over unitary matrices. Assume that the number of nonzero elements in ρ is only polynomial in n. For any given failure probability $0 < \delta < 1$ and error tolerance $\gamma > 0$, consider a collection of unitary matrices $\{U_i\}_{i=1}^m$, where $m = poly(n)$ with coefficient depending on δ and γ, sampled independently from distribution $\mathcal{D}$, with the corresponding probability distributions $h(U_i) = diag(U_i \rho U_i^\dagger)$. Suppose there exists a learning model f from a sufficiently expressive hypothesis space $\mathcal{F}$ such that:*

(1) f corresponds to some quantum state σ such that $f(U_i) = diag(U_i \sigma U_i^\dagger)$ for all $i = 1, \ldots, m$,

(2) f achieves empirical error bound such that $\frac{1}{m}\sum_{i=1}^{m}\ell(h(U_i), f(U_i)) \leq \eta$ for a given $\eta > 0$.

Then with probability at least $1-\delta$, the expected loss satisfies:

$$\mathbb{E}_{\mathcal{D}}\,\ell(h(U), f(U)) \leq \gamma + 2\eta.$$

The theorem requires three assumptions. First, the number of nonzero elements in the matrix should be polynomial qubit number, which enables dimension reduction of the feature map in Section 4. Second, the output of the learning algorithm should be consistent with measurement results from some quantum state, allowing us to express the algorithm's output as a functional of feature functions. In Section 5, for neural network, we achieve this requirement through a penalty term in the loss function. Third, the loss function should remain below a certain threshold, which is achievable for algorithms with sufficient learning capacity.

For the proof, we diverge from the method employed in [21], which relies on amplification techniques and adaptive protocols for iterative density matrix reconstruction. Instead, we follow the theoretical framework established in [15], which offers a more direct approach to estimate measurement statistics without requiring complete state reconstruction. In particular, we employ local Rademacher complexity, a fundamental concept in statistical learning theory [18], to establish our proof. Our proof strategy focuses on controlling the local Rademacher complexity of the loss function. We first decompose the local Rademacher complexity of the loss function into the local Rademacher complexities of its components. Next we construct a kernel function and the corresponding feature map such that the components can be expressed as functionals of the feature map. We then estimate the expected loss over the entire distribution using empirical loss and local Rademacher complexity. Finally, we employ random projection techniques to reduce the required dataset size to polynomial scale.

3. Rademacher Complexity

In order to prove Theorem 1, we need to bound the actual error of the model, which requires measuring the learning capacity through the Rademacher complexity. This useful tool helps us measure how effectively a learning algorithm can capture patterns in random data. The definition of Rademacher complexity is given below.

Definition 1 (Rademacher Complexity)**.** *Let $(\mathcal{X}, P)$ be a probability space and $\mathcal{F}$ be a set of measurable functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Given a positive integer m, consider:*

(1) m i.i.d. samples from $(\mathcal{X}, P)$, denoted by $X_1, X_2, \ldots, X_m$,

(2) m i.i.d. Rademacher variables with $\Pr(\sigma_i = \pm 1) = \frac{1}{2}$, denoted by $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$.

For any $f \in \mathcal{F}$, we define:

$$\begin{cases} P_m f = \frac{1}{m}\sum_{i=1}^{m} f(X_i), & \textit{(empirical mean)} \\ Pf = \mathbb{E}f(X), & \textit{(expected value)} \\ R_m f = \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(X_i). & \textit{(Rademacher average)} \end{cases}$$

The empirical Rademacher complexity of $\mathcal{F}$ is then defined as:

$$\mathbb{E}_\sigma R_m \mathcal{F} = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(X_i)\middle|(X_1, X_2, \ldots, X_m)\right],$$

where $R_m\mathcal{F} = \sup_{f\in\mathcal{F}} R_m f$.

The local Rademacher complexity (LRC) is defined in the same way but with an additional constraint $P_m f^2 < r$, which is denoted as $\mathbb{E}_\sigma R_m \mathcal{F}(r)$. The LRC serves as a fundamental tool in theoretical machine learning proofs, especially its sub-root property [18].

Definition 2. *A function $\phi : [0, +\infty) \to \mathbb{R}$ is called sub-root if it satisfies the following conditions:*

(1) ϕ is non-negative,

(2) ϕ is non-decreasing,

(3) $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$.

This sub-root property ensures the existence of a unique fixed point for ϕ, which is shown in the following lemma.

Lemma 1. *A sub-root function ϕ, which is not identically zero, possesses exactly one positive fixed point, denoted as r^*. For any $r > 0$, the inequality $\phi(r) \leq r$ holds if and only if $r^* \leq r$. Furthermore, given two sub-root functions ϕ_1 and ϕ_2 where $\phi_1(r) < \phi_2(r)$ for all $r > 0$, the fixed point r_1^* of ϕ_1 must be smaller than the fixed point r_2^* of ϕ_2.*

Proof. The proof begins by establishing the continuity of ϕ. For any $r_1 > r_2 > 0$, the non-decreasing property of ϕ implies $|\phi(r_1) - \phi(r_2)| = \phi(r_1) - \phi(r_2)$. Given that $\phi(r)/\sqrt{r}$ is non-increasing, we can derive $\phi(r_1)/\sqrt{r_2} \leq \sqrt{r_1}\phi(r_2)/r_2$, leading to the inequality:

$$\phi(r_1) - \phi(r_2) \leq \phi(r_2)\frac{\sqrt{r_1} - \sqrt{r_2}}{\sqrt{r_2}}.$$

The continuity of ϕ follows as $|\phi(r_1) - \phi(r_2)|$ approaches zero when r_1 approaches r_2 from either direction. The function $\phi(r)/r$ inherits continuity in $(0, +\infty)$ and maintains non-negativity. The strict monotonic decrease of $1/\sqrt{r}$ in $(0, +\infty)$ ensures that $\phi(r)/r$ is also strictly decreasing.

If this ratio $\phi(r)/r$ consistently exceeds 1, then $\lim_{r\to+\infty} \phi(r)/\sqrt{r}$ would diverge to infinity, which contradicts the sub-root property (3). Conversely, if $\phi(r)/r$ is consistently less than 1, then $\lim_{r\to 0^+} \phi(r)/\sqrt{r} = 0$ would imply that $\phi(r)$ vanishes in $[0, +\infty)$, contradicting the non-triviality of ϕ. Therefore, equation $\phi(r)/r = 1$ must have exactly one positive solution, guaranteed by monotonicity. When $\phi(r) \leq r$ for some $r > 0$, we can deduce that $\phi(t)/t \leq 1$ for all $t \geq r$, implying $r^* \leq r$. The converse follows analogously.

Finally, for two sub-root functions ϕ_1 and ϕ_2, where $\phi_1(r) < \phi_2(r)$ for all $r > 0$, we can conclude that $\phi_1(r_2^*) < \phi_2(r_2^*) = r_2^*$, which implies $r_1^* < r_2^*$. □

Lemma 1 establishes a fundamental property of sub-root functions. It shows that a sub-root function has a unique fixed point. It also provides us with a method to bound this fixed point. It turns out that the LRC has a sub-root nature, as given in the following lemma [18].

Lemma 2. *Let $\mathcal{F}$ be a class of measurable functions. Consider a functional $T : \mathcal{F} \to [0, +\infty)$ that satisfies $T(\alpha f) \leq \alpha^2 T(f)$ for all $f \in \mathcal{F}$ and $\alpha \in [0, 1]$. Given $\hat{f} \in \mathcal{F}$, define a random function ϕ as LRC on a subset of $\mathcal{F}$:*

$$\phi(r) = \mathbb{E}_\sigma R_m\{f \in \mathcal{F} : T(f - \hat{f}) \leq r\},\ r \in [0, +\infty).$$

Then both ϕ and its expectation $\mathbb{E}\phi(r)$ are sub-root functions.

Now consider a supervised learning framework with input space $\mathcal{X}$ and output space $\mathcal{Y}$, where the joint distribution P is defined on $\mathcal{X} \times \mathcal{Y}$. Given a training sample $\{(X_i, Y_i)\}_{i=1}^m$ drawn independently from P, the objective is to identify a function $f : \mathcal{X} \to \mathcal{Y}$ from a function class $\mathcal{F}$ that minimizes the expected loss

$$\mathbb{E}\ell_f = \mathbb{E}\ell(f(X), Y),$$

where ℓ is an arbitrary loss function on $\mathcal{Y} \times \mathcal{Y}$. Let $\ell_\mathcal{F}$ denote the loss class induced by $\mathcal{F}$, formally defined as:

$$\ell_\mathcal{F} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}.$$

Define a star-hull around zero as:

$$star(\ell_\mathcal{F}, 0) = \{\alpha f : f \in \ell_\mathcal{F}, \alpha \in [0, 1]\}.$$

Building upon Lemma 2, one can establish the following theorem [18]:

Theorem 2. *Consider a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$. Given $x > 0$, define the function $\hat{\phi}_{m,x}$ as:*

$$\hat{\phi}_{m,x}(r) = 20\mathbb{E}_\sigma R_m\{f \in star(\ell_\mathcal{F}, 0) : P_m f^2 \leq 2r\} + \frac{13x}{m}. \tag{1}$$

Let $\hat{r}^*$ be the fixed point of $\hat{\phi}_{m,x}$. Then for any $K > 1$, the following inequality holds for all functions $f \in \mathcal{F}$:

$$P\ell_f \leq \frac{K}{K-1} P_m \ell_f + 6K\hat{r}^* + \frac{x(11+5K)}{m} \tag{2}$$

with probability at least $1 - 3e^{-x}$.

This theorem establishes a relationship between the empirical error and the actual error through local Rademacher complexity. When the Rademacher term on the right-hand side of equation (2) can be sufficiently small and the fixed point $\hat{r}^*$ approaches the origin, it is expected that the actual error $P\ell_f$ approaches zero. To quantify this convergence, an estimation of the Rademacher complexity is required. This estimation is based on two results, the first is related to the Rademacher complexity of vector-valued functions [22].

Theorem 3. *For any set $\mathcal{X}$ and a sequence $(X_1, \ldots, X_m) \in \mathcal{X}^m$, consider a class of functions $\mathcal{F}$ mapping from $\mathcal{X}$ to ℓ_2 space. Given a collection of functions $h_i : \ell_2 \to \mathbb{R}$ with Lipschitz constant L, the following inequality holds:*

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i h_i(f(X_i)) \leq \sqrt{2} L \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sum_{k} \sigma_{ik} f_k(X_i), \tag{3}$$

where σ_{ik} represents the independent doubly indexed Rademacher sequence and $f_k(X_i)$ denotes the k-th component of the vector $f(X_i)$.

Theorem 3 establishes a fundamental relationship between the Rademacher complexity of vector-valued functions and their scalar components. In particular, it provides a practical way for complexity estimation by showing that the complexity of a composite vector-valued function can be bounded by the complexities of its components. To obtain a more accurate complexity bound, we adapt the theorem from Cortes et al. [23] with an additional norm scaling factor.

Theorem 4. *Let $k(x, \tilde{x})$ be a Mercer kernel (symmetric and positive semidefinite) with its associated feature map Φ_k. Consider its eigenvalue decomposition:*

$$k(x, \tilde{x}) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)^T \varphi_j(\tilde{x}),$$

where $(\lambda_j)_{j=1}^{\infty}$ represents the sequence of eigenvalues in descending order with corresponding eigenfunctions $(\varphi_j)_{j=1}^{\infty}$. Given $B > 0$, define the function class:

$$\mathcal{F} = \{f_w = (x \mapsto \langle \omega, \Phi_k(x) \rangle) : \|w\|_2 \leq B\}.$$

Then, for any $r > 0$, it holds the following bound on the expected Rademacher complexity:

$$\mathbb{E}_\sigma R_m \mathcal{F}(r) \leq B \sqrt{\frac{2}{m} \min_{\theta \geq 0} \left(\theta r + \sum_{j > \theta} \lambda_j \right)} = B \sqrt{\frac{2}{m} \sum_{j=1}^{\infty} \min(r, \lambda_j)}. \tag{4}$$

Theorem 4 shows that if we know the eigenvalues of the kernel function, we can estimate the local Rademacher complexity of functionals of its feature map. The construction of such a feature map will be presented in Section 4.

4. Polynomial Complexity

In the preceding sections, we introduced the local Rademacher complexity and related theorems. This section focuses on proving Theorem 1. We begin by establishing the Lipschitz continuity of the loss function class, which is

needed in Theorem 3. In the subsequent analysis, $\mathcal{G}$ represents the function class that maps the input space $\mathcal{X}$ to the distribution space $\mathbb{R}^N$, where $N = 2^n$ is the dimension of n-qubit Hilbert space.

Lemma 3. *For a given distribution function $\bar{f} \in \mathcal{F}$, let us define a new function class $\mathcal{G}_{\bar{f}}$ as:*

$$\mathcal{G}_{\bar{f}} = \{f(x) - \bar{f}(x) : f \in \mathcal{F}\}.$$

Consider the function $\ell_{\mathcal{G}} : \mathcal{G}_{\bar{f}} \to \mathbb{R}$ defined by:

$$\ell_{\mathcal{G}}(g(x)) = \sum_{k=1}^{N} (g_k(x))^2, \quad g \in \mathcal{G}_{\bar{f}},$$

where $g_k(x)$ is the k-th component of $g(x)$. Then every function in $star(\ell_{\mathcal{G}}, 0)$ is Lipschitz continuous with Lipschitz constant 4.

Proof. It is sufficient to prove

$$|\ell_{\mathcal{G}}(g(x)) - \ell_{\mathcal{G}}(h(x))| \leq 4\|g(x) - h(x)\|, \tag{5}$$

where $g = f_1 - \bar{f}$ and $h = f_2 - \bar{f}$ both belong to $\mathcal{G}_{\bar{f}}$.

Let us expand the left-hand side of equation (5):

$$\begin{aligned}
& |\ell_{\mathcal{G}}(g(x)) - \ell_{\mathcal{G}}(h(x))| \\
= \quad & \left| \sum_{i=1}^{N} (g_i(x))^2 - \sum_{i=1}^{N} (h_i(x))^2 \right| \\
= \quad & \left| \sum_{i=1}^{N} (g_i(x) - h_i(x))(g_i(x) + h_i(x)) \right| \\
\leq \quad & \left| \sum_{i=1}^{N} f_{1,i}(x)(g_i(x) - h_i(x)) \right| + \left| \sum_{i=1}^{N} f_{2,i}(x)(g_i(x) - h_i(x)) \right| + \left| 2 \sum_{i=1}^{N} \bar{f}_i(x)(g_i(x) - h_i(x)) \right| \\
\leq \quad & \|g(x) - h(x)\|_\infty \left(\sum_{i=1}^{N} f_{1,i}(x) + \sum_{i=1}^{N} f_{2,i}(x) + 2 \sum_{i=1}^{N} \bar{f}_i(x) \right) \\
\leq \quad & 4 \sqrt{\sum_{i=1}^{N} (g_i(x) - h_i(x))^2} \\
= \quad & 4\|g - h\|_2.
\end{aligned}$$

The second inequality uses the fact that all $f_{1,i}(x), f_{2,i}(x), \bar{f}_i(x)$ are components of probability distributions. □

We then introduce a specific feature map for unitary matrices that plays a crucial role in our estimation. Let $SU(N)$ denote the special unitary group of degree $N = 2^n$, consisting of all $N \times N$ unitary matrices with determinant equal to 1. For a unitary matrix $U := (u_{ij}) \in SU(N)$ and $1 \leq k \leq N$, we define the feature map $\Phi^{(k)}(U) = \{(u_{ki}\overline{u_{kj}})_{(i,j)},\ i, j = 1, \ldots, N\} \in \mathbb{C}^{N^2}$. Given any two unitary matrices U_1, U_2, we define the corresponding kernel function as their inner product in the feature space:

$$k(U_1, U_2) = \langle \Phi^{(k)}(U_1), \Phi^{(k)}(U_2) \rangle.$$

By Theorem 4, this kernel admits a spectral decomposition:

$$k(U_1, U_2) = \sum_{i=1}^{K} \lambda_i \varphi_i(U_1) \varphi_i^T(U_2),$$

where K represents the rank of the kernel matrix, bounded above by N^2. The construction of this feature map naturally leads to the following result.

Theorem 5. *Let ρ be a density matrix of order $N = 2^n$ and $U \in SU(N)$. Then each element of the vector $h = diag(U\rho U^\dagger)$ can be expressed as an inner product between the feature map and a vector whose ℓ_2-norm is strictly bounded by 1.*

Proof. For $k = 1 \ldots, N$, we have

$$h_k(U) = \sum_{i,j=1}^{N} u_{ki}\overline{u_{kj}}\rho_{ij}.$$

Let $w = (\rho_{ij}) \in \mathbb{C}^{N^2}$. Then we can express $h_k(U)$ as the inner product:

$$h_k(U) = \langle w, \Phi^{(k)}(U)\rangle.$$

To establish the norm bound, observe that $\|w\|_2$ equals the Frobenius norm of ρ. Let $\{\sigma_i\}_{i=1}^N$ denote the singular values of ρ. By the properties of density matrices, it holds $\sum_{i=1}^{N} \sigma_i = \mathrm{tr}\rho = 1$. Therefore:

$$\|\rho\|_F = \sqrt{\sum_{i=1}^{N} \sigma_i^2} \leq \sqrt{\sum_{i=1}^{N} \sigma_i} = 1,$$

where the inequality follows from the fact that $0 \leq \sigma_i \leq 1$. This establishes the conclusion that $\|w\|_2 \leq 1$. □

We also require the Johnson-Lindenstrauss (JL) lemma [24], which is fundamental to the dimensional reduction analysis.

Theorem 6 (Johnson-Lindenstrauss Lemma)**.** *For any sample size $m \geq 1$, any finite point set $\mathcal{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^m$, and any error tolerance $\varepsilon \in (0, 1)$, let q be a positive integer satisfying:*

$$q \geq \frac{C \ln m}{\varepsilon^2(1-\varepsilon)},$$

where C is a positive constant that only depends on d. Then there exists a linear map $F : \mathbb{R}^d \to \mathbb{R}^q$ of the form $F(x) = V^T x$, where $V = (v_{ij}) \in \mathbb{R}^{d\times q}$ and v_{ij} are i.i.d. random variables with zero mean and scaled unit variance, such that it holds:

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|F(x_i) - F(x_j)\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2 \tag{6}$$

for all pairs $x_i, x_j \in \mathcal{X}$ with probability at least $1 - \delta$, where $\delta = 2e^{-(\varepsilon^2(1-\varepsilon))(q/4)}$.

Remark 1. *A canonical example is simply choosing $v_{ij} \sim \mathcal{N}(0, 1/q)$.*

Now we have all the tools necessary to establish our main result. We begin by presenting a weaker variant of Theorem 1:

Theorem 7. *Let ρ be an n-qubit quantum state and $\mathcal{D}$ be a distribution over unitary matrices. For any given failure probability $0 < \delta < 1$ and error tolerance $\gamma > 0$, consider a collection of unitary matrices $\{U_i\}_{i=1}^m$, where $m = poly(2^n)$ with coefficient depending on δ and γ, sampled independently from distribution $\mathcal{D}$, with the corresponding probability distributions $h(U_i) = diag(U_i\rho U_i^\dagger)$. Suppose there exists a learning model f from a sufficiently expressive hypothesis space $\mathcal{F}$ that:*

(1) f corresponds to some quantum state σ such that $f(U_i) = diag(U_i\sigma U_i^\dagger)$ for all $i = 1, \ldots, m$,

(2) f achieves empirical error bound such that $\frac{1}{m}\sum_{i=1}^{m} \ell(h(U_i), f(U_i)) \leq \eta$ for a given $\eta > 0$.

Then with probability at least $1 - \delta$, the expected loss satisfies:

$$\mathbb{E}_{\mathcal{D}}\, \ell(h(U), f(U)) \leq \gamma + 2\eta.$$

Remark 2. *Theorems 1 and 7 differ only in their sample complexity: $m = poly(n)$ in Theorem 1 versus $m = poly(2^n)$ in Theorem 7.*

Proof. According to the assumption, we can represent each probability distribution $h(U_i)$ with $i = 1, \ldots, m$ as an N-dimensional vector, where $N = 2^n$ corresponds to the dimension of the n-qubit Hilbert space. Let $h_k(U_i)$ denote the probability of measuring the system in the k-th computational basis state of $h(U_i)$, where the states are encoded in binary representation, i.e., $|0\rangle^{\otimes n}$ corresponds to index 0, $|0\rangle^{\otimes(n-1)}|1\rangle$ corresponds to index 1, and so forth.

We denote the unitary matrices U_i as input variables x_i. According to Theorem 3 and Lemma 3, it holds:

$$\begin{aligned}
\mathbb{E}_\sigma R_m \ell_{\mathcal{G}} &= \frac{1}{m}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^{m} \sigma_i \ell(g(x_i), h(x_i)) \\
&\leq \frac{4\sqrt{2}}{m}\mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^{m}\sum_{k=1}^{N} \sigma_{ik}(g_k(x_i) - h_k(x_i)) \\
&\leq \frac{4\sqrt{2}}{m}\sum_{k=1}^{N} \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^{m} \sigma_{ik}(g_k(x_i) - h_k(x_i)) \\
&= 4\sqrt{2}\sum_{k=1}^{N} \mathbb{E}_\sigma R_m \mathcal{G}_k,
\end{aligned} \tag{7}$$

where g_k is the k-th component of g, and $\mathcal{G}_k = \{g_k - h_k : g \in \mathcal{G}\}$.

Equation (7) enables us to estimate the Rademacher complexity of the loss function through its component-wise Rademacher complexities. Given $r > 0$, considering the constraint

$$\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{k=1}^{N}(g_k(x_i) - h_k(x_i))^2\right)^2 \leq r,$$

we can deduce that $\left(\sum_{k=1}^{N}(g_k(x_i) - h_k(x_i))^2\right)^2 \leq mr$ for all $i = 1, \ldots, m$. This implies

$$\sum_{k=1}^{N}(g_k(x_i) - h_k(x_i))^2 \leq \sqrt{mr}.$$

Summing over i yields $\sum_{i=1}^{m}\sum_{k=1}^{N}(g_k(x_i) - h_k(x_i))^2 \leq m\sqrt{mr}$, or equivalently,

$$\sum_{k=1}^{N}\frac{1}{m}\sum_{i=1}^{m}(g_k(x_i) - h_k(x_i))^2 \leq \sqrt{mr}. \tag{8}$$

Denote $a_k = \frac{1}{m}\sum_{i=1}^{m}(g_k(x_i) - h_k(x_i))^2$. For any interval $[\sqrt{mr}/2^j, \sqrt{mr}/2^{j-1}]$, the number of a_k within this interval cannot exceed 2^j, where $1 \leq j \leq n$. For $j = n$, we extend the lower bound to 0 so that $0 \leq a_k \leq \sqrt{mr}/2^{n-1}$. Consequently, from equation (7), it implies

$$\mathbb{E}_\sigma R_m \ell_{\mathcal{G}}(r) \leq 4\sqrt{2}\sum_{j=1}^{n} 2^j \mathbb{E}_\sigma R_m \mathcal{G}_k\left(\frac{\sqrt{rm}}{2^{j-1}}\right). \tag{9}$$

Based on Theorem 5, we can get that $g_k = w \cdot \Phi(U)$ and $h_k = v \cdot \Phi(U)$, which implies

$$\|w - v\| \leq \|w\| + \|v\| \leq 2.$$

From Theorem 4, we obtain

$$\mathbb{E}_\sigma R_m \mathcal{G}_k\left(\frac{\sqrt{rm}}{2^{j-1}}\right) \leq 2\sqrt{\frac{2}{m}\sum_{k=1}^{K}\min\left(\frac{\sqrt{rm}}{2^{j-1}}, \lambda_k\right)},$$

where $K \leq N^2$. Since k ranges from 1 to K, this yields

$$\mathbb{E}_\sigma R_m \mathcal{G}_k\left(\frac{\sqrt{rm}}{2^{j-1}}\right) \leq 2\sqrt{\frac{2K}{m}\frac{\sqrt{rm}}{2^{j-1}}}. \tag{10}$$

Substituting equation (10) into equation (9) produces

$$\mathbb{E}_\sigma R_m \ell_{\mathcal{G}}(r) \leq 8\sqrt{2} r^{\frac{1}{4}} m^{-\frac{1}{4}} \sqrt{K} \sum_{j=1}^{n} 2^{\frac{j}{2}} \leq 32\sqrt{2}\sqrt{NK} m^{-\frac{1}{4}} r^{\frac{1}{4}}. \tag{11}$$

Combining equation (11) and equation (1) in Theorem 2, we get that

$$\hat{\phi}_{m,x}(r) \leq 1280 r^{\frac{1}{4}} m^{-\frac{1}{4}} N^{\frac{3}{2}} + \frac{13x}{m} = \tilde{\phi}_{m,x}(r). \tag{12}$$

Based on Lemma 1, we now aim to find the upper bound of the fixed point $\tilde{r}$ of $\tilde{\phi}_{m,x}(r)$, where $\tilde{r}$ satisfies the equation

$$1280 \tilde{r}^{\frac{1}{4}} m^{-\frac{1}{4}} N^{\frac{3}{2}} + \frac{13x}{m} = \tilde{r}.$$

This equation can be rewritten as $\tilde{r}^{\frac{1}{4}}(\tilde{r}^{\frac{3}{4}} - a) = b$, where $a = 1280 N^{\frac{3}{2}} m^{-\frac{1}{4}}$ and $b = \frac{13x}{m}$. Let

$$\hat{r} = \left(1280 N^{\frac{3}{2}} m^{-\frac{1}{4}} + \frac{1}{\sqrt{m}}\right)^{\frac{4}{3}} + \left(\frac{13x}{\sqrt{m}}\right)^4. \tag{13}$$

Then $\hat{r}$ satisfies $\hat{r}^{\frac{3}{4}} \geq a + \frac{1}{\sqrt{m}}$ and $\hat{r}^{\frac{1}{4}} \geq \sqrt{m} b$, so using Lemma 1, we get $\hat{r} \geq \tilde{r}$.

Applying Theorem 2, we can establish:

$$\begin{aligned} P\ell_f \leq & \frac{K}{K-1} P_m \ell_f + 6K\hat{r} + \frac{x(11+5K)}{m} \\ \leq & \frac{K}{K-1} P_m \ell_f + 6K\left(\left(1280 N^{\frac{3}{2}} m^{-\frac{1}{4}} + \frac{1}{\sqrt{m}}\right)^{\frac{4}{3}} + \left(\frac{13x}{\sqrt{m}}\right)^4\right) + \frac{x(11+5K)}{m} \end{aligned} \tag{14}$$

with probability $1 - 3e^{-x}$. For any $0 < \delta < 1$, setting $K = 2$ and $x = -\ln\left(\frac{\delta}{3}\right)$, we can select an appropriate $m = poly(2^n)$, to ensure the sum of the last two terms remains below a given error tolerance γ. In particular, it holds $m = \mathcal{O}\left(N^{\frac{8}{3}} \left(\ln\left(\frac{\delta}{3}\right)\right)^2 \frac{1}{\gamma^3}\right)$ from equation (14). □

While our analysis shows that quantum state reconstruction is possible, the exponential scaling of measurement requirements $m = ploy(2^n)$ remains a fundamental challenge for practical applications. To address this limitation, we extend our analysis to Theorem 1 by incorporating sparsity constraints on the density matrix representation of quantum states, which demonstrates a substantial reduction in measurement complexity.

Proof. of Theorem 1: Here we follow the notations established in Theorem 7. From Theorem 7, for random samples $\{(x_i, y_i)\}_{i=1}^m$ drawn from distribution $\mathcal{D}$, where $m = poly(2^n)$, we obtain equation (14). Let $\widetilde{\mathcal{D}}$ denote the uniform distribution over these samples.

Choose a constant $\varepsilon = \min\{\frac{\gamma}{2}, \frac{1}{2}\}$. Following Lemma 6, we define $q = \frac{C_1 n}{\varepsilon^2(1-\varepsilon)} \geq \frac{C \ln(2m)}{\varepsilon^2(1-\varepsilon)}$, where $C > 0, C_1 > 0$ are constants. We construct a random projection matrix $V \in \mathbb{R}^{2m \times q}$ with entries independently sampled from $\mathcal{N}(0, \frac{1}{q})$. With probability $1 - 2(2m)^{-\frac{C}{4}}$, it holds

$$(1-\varepsilon)\|z_i - z_j\| \leq \|Vz_i - Vz_j\| \leq (1+\varepsilon)\|z_i - z_j\| \tag{15}$$

for all pairs $z_i, z_j \in \{y_1, \ldots, y_m, f(x_1), \ldots, f(x_m)\}$. Let $\tilde{f}(x) = Vf(x)$ and $\tilde{y} = Vy$. From equation (15), we derive:

$$(1-\varepsilon)^2 \ell(f, y) \leq \ell(\tilde{f}, \tilde{y}) \leq (1+\varepsilon)^2 \ell(f, y). \tag{16}$$

Following the proof structure of Theorem 7 with distribution $\widetilde{\mathcal{D}}$, we now try to determine the sample size $\widetilde{m}$. Given the compression to dimension q, equation (7) becomes:

$$\mathbb{E}_\sigma R_{\widetilde{m}} \ell_{\mathcal{G}} \leq 4\sqrt{2} \sum_{k=1}^{q} \mathbb{E}_\sigma R_{\widetilde{m}} \mathcal{G}_k.$$

Using the sparsity hypothesis, we can reduce the dimension of the feature map $\Phi^{(k)}$, thus reducing the dimensionality factor K in equation (10) to M (the number of nonzero elements in ρ, which is $poly(n)$). This yields the refined bound:

$$r^* \leq \left(1280\sqrt{Mq} \cdot \widetilde{m}^{-\frac{1}{4}} + \frac{1}{\sqrt{\widetilde{m}}}\right)^{\frac{4}{3}} + \left(\frac{13x}{\sqrt{\widetilde{m}}}\right)^4.$$

Choosing $\widetilde{m} = \mathcal{O}\left((Mq)^{\frac{4}{3}} \left(\ln\left(\frac{\delta}{3}\right)\right)^2 \frac{1}{\tilde{\gamma}^3}\right) = poly(n)$ ensures:

$$P\ell_{\tilde{f}} \leq 2P_{\widetilde{m}}\ell_{\tilde{f}} + \tilde{\gamma},$$

where $0 < \tilde{\gamma} < 1$. Since $\widetilde{\mathcal{D}}$ is discrete, we have $P\ell_{\tilde{f}} = P_m\ell_{\tilde{f}}$. The desired result follows from equation (14) and equation (16). □

5. Recurrent Neural Network (RNN)

In this section, we address the practical implementation of Theorem 1. A challenge lies in determining an efficient parameterization of unitary matrices that can serve as viable input to our model. Since the dimension of the unitary group $SU(2^n)$ grows exponentially with n, direct parameterization of arbitrary unitary matrices becomes impractical. Therefore, for practical implementation, we restrict our attention to a subset of unitary matrices that can be decomposed as Kronecker products of 2×2 unitary matrices:

$$U = U_1 \otimes U_2 \otimes \cdots \otimes U_n, \quad U_i \in SU(2).$$

This reduction to local operations simplifies our approach by focusing on the parameterization of $SU(2)$ matrices. We express the unitary matrices as:

$$\begin{pmatrix} \cos\frac{\theta}{2} e^{i(\phi+\psi)/2} & \sin\frac{\theta}{2} e^{i(\phi-\psi)/2} \\ -\sin\frac{\theta}{2} e^{i(-\phi+\psi)/2} & \cos\frac{\theta}{2} e^{i(-\phi-\psi)/2} \end{pmatrix},$$

where θ, ϕ, ψ are all from 0 to 2π. The volume form for this parameterized unitary matrix is proportional to $d\cos(\theta) d\phi d\psi$. While our analysis allows for arbitrary probability distributions of unitary matrices, we adopt the Haar measure for sampling to ensure a uniform error distribution across randomly sampled unitary matrices. This choice leads to the following parameterization:

$$\begin{cases} \theta &= \arccos\theta', \\ \phi &= \phi', \\ \psi &= \psi', \end{cases}$$

where ϕ', ψ' are sampled uniformly from the original defined interval, and θ' is sampled uniformly in $[-1, 1]$.

To validate Theorem 1, we design a series of numerical experiments using the Long Short-Term Memory (LSTM) as our main statistical learning model. The LSTM model, known for its capability to process sequential data with long-term dependencies, is particularly suitable for our application. In our framework, each measurement outcome can be represented as a sequence of binary digits (0 or 1), analogous to a sentence in natural language processing. The model generates conditional probabilities for each subsequent measurement outcome based on current and previous results. The joint probability of the entire measurement sequence is then computed as the product of these conditional probabilities.

The prediction phase follows a different process from the training phase. During prediction, the neural network accepts specified conditions as input and generates conditional probabilities at each sequential step. These probabilities guide a sampling process where each binary outcome ("word") is selected according to the computed probability distribution. The selected outcome is then fed as input to the subsequent step. The final probability of the complete

quantum state is computed as the product of these sequential conditional probabilities. This prediction mechanism is depicted in Figure 1.

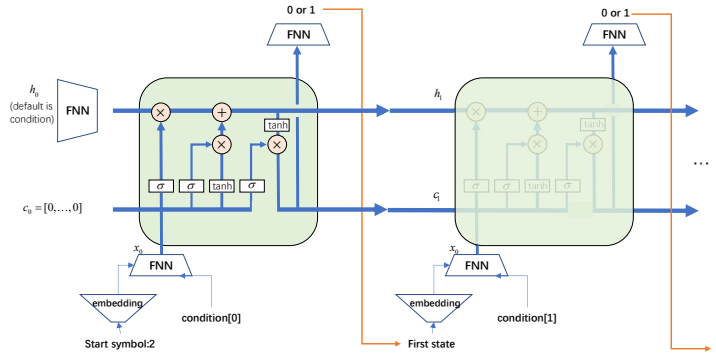

Figure 1. The diagram illustrates the RNN architecture for sampling. FNN represents feedforward neural network. The output's 0-1 distribution allows for sampling a specific state, which is then passed on to the next step as a measurement result.

To enhance the predictive capability, we incorporate quantum measurement context through a dual-input structure: (1) the measurement outcome of the previous particle, and (2) the parameterization of the unitary operator U corresponding to the current particle's measurement basis. This design enables the network to learn the correlation between measurement outcomes and their corresponding basis transformations, thereby capturing the quantum state's behavior under different measurement bases. For sequence initialization, we employ a designated token '2', which appears exclusively at the beginning of each measurement sequence. This initialization token provides a clear way of sequence boundaries in the training data.

We now introduce the loss function employed in this work. We adopt the cross-entropy loss as our objective function due to its effectiveness in distribution approximation tasks. The cross-entropy loss between the true distribution p and the predicted distribution q is defined as:

$$\mathcal{L}_{cross-entropy}(p,q) = \sum_i p_i \log \frac{1}{q_i}.$$

Since our training samples are drawn independently from the true distribution p, by the law of large numbers, the empirical loss converges to:

$$\mathcal{L}_{cross-entropy}(p,q) \approx \mathbb{E}_{x\sim p}\left[\log \frac{1}{q(x)}\right] = \sum_i \log \frac{1}{q_i}.$$

This formulation eliminates the need to compute the true probabilities p_i, thereby reducing memory requirements.

To ensure that our neural network correctly represents the quantum state transformation $h(U) = \mathrm{diag}(U\rho U^\dagger)$, we incorporate two constraints into the loss function. The first is a parameterization invariance constraint: parameter sets $\{\theta, \phi, \psi\}$ and $\{-\theta, \phi + \pi, \psi + \pi\}$ must produce identical outputs, as they represent the same unitary transformation. The second is the constraint derived from the properties of $f(U) = \mathrm{diag}(U\rho U^\dagger)$. For the single-qubit case, these properties are derived as follows and they can naturally extend to multi-qubit systems through tensor products. Consider the following matrix representations:

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix}, U = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}, U_1 = \begin{pmatrix} u_{11} & -u_{12} \\ -u_{21} & u_{22} \end{pmatrix}.$$

This yields the functional form:

$$\begin{aligned} f(U) \quad = \quad & (\rho_{11}|u_{11}|^2 + 2\mathrm{Re}(\rho_{12}\overline{u_{12}}u_{11}) + \rho_{22}|u_{12}|^2, \\ & \rho_{11}|u_{21}|^2 + 2\mathrm{Re}(\rho_{12}\overline{u_{21}}u_{22}) + \rho_{22}|u_{22}|^2)^T. \end{aligned}$$

Under the chosen parameterization, we derive a symmetric relation:

$$f(U) + f(U_1) = 2\left(\cos^2\left(\frac{\theta}{2}\right) f(I) + \sin^2\left(\frac{\theta}{2}\right) f(X)\right),$$

where

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Furthermore, let

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, H_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix},$$

and $\bar{f}_1 = \left(f(H_1) - \frac{f(I)+f(X)}{2}\right), \bar{f}_2 = \left(f(H_2) - \frac{f(I)+f(X)}{2}\right)$, we obtain the antisymmetric component:

$$f(U) - f(U_1) = 2\big(\sin(\theta)\cos(\psi)\bar{f}_1 - \sin(\theta)\sin(\psi)\bar{f}_2\big).$$

These two constraints are incorporated into the total loss function via l_1 regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cross-entropy}} + \lambda(\mathcal{L}_{\text{symm}} + \mathcal{L}_{\text{invariant}}),$$

where λ serves as the regularization coefficient.

We have also investigated replacing the LSTM model with a multi-head attention mechanism, which has shown remarkable success in various sequence modeling tasks. However, LSTM maintains a higher accuracy in our case and offers easier training procedures. This observation suggests that the sequential nature of quantum measurements may be better modeled by the LSTM's memory structure, which is demonstrated in the following section.

6. Numerical Experiments

Before proceeding with the experimental analysis, we detail the configuration of our numerical experiments. To quantify the performance of the model, we employ the classical fidelity measure between the predicted distribution p and the true distribution q:

$$f(p,q) = \sum_{i=1}^{N} \sqrt{p_i q_i}.$$

This metric provides a natural measure of the similarity between the probability distributions of the measurement.

We now detail the hyperparameters employed in our implementation. The model was trained for 50 epochs across all experiments. While this choice lacks rigorous foundation, empirical evidence shows its consistency in achieving convergence across various system sizes. The regularization coefficient λ was scaled dynamically with the particle number, ranging from 5 to 50. For smaller systems, we found that a smaller λ value leads to faster loss convergence, while larger systems benefit from increased λ values to enforce stronger constraints on the neural network's behavior.

To address the computational complexity of evaluating $\mathcal{L}_{\text{symm}}$ and $\mathcal{L}_{\text{invariant}}$, we implement a stochastic index sampling strategy. Rather than computing losses over all indices, we randomly select 50 indices for loss calculation in each iteration. This approach reduces computational overhead while maintaining the effectiveness of the constraints. To ensure polynomial scaling of computational resources, we impose strategic constraints on the model architecture. The dimensionality of both the RNN hidden states and the word vector embeddings is set to $2n^2$, where n denotes qubit number.

6.1. TFIM Models

For initial validation, we test our model against the ground state of the transverse field Ising model (TFIM), a system for studying quantum phase transitions. The TFIM Hamiltonian is expressed as:

$$H = J_z \sum_{\langle i,j \rangle} \sigma_i^z \sigma_j^z - J_x \sum_i \sigma_i^x,$$

where $\langle i,j \rangle$ denotes nearest-neighbor interactions, σ_i^z and σ_i^x are Pauli operators acting on site i. We fix the ferromagnetic coupling strength $J_z = 1$ as our energy scale and vary the transverse field strength $J_x \in [0,2]$ to probe different

quantum phases. This parameterization enables us to investigate the system's behavior at the quantum critical point at $J_x / J_z = 1$.

The model's performance, measured by fidelity and illustrated in Figure 2, consistently exceeds 90% across different parameter regimes. These results demonstrate the model's robustness and its capability to capture essential features of the quantum state's probability distribution.

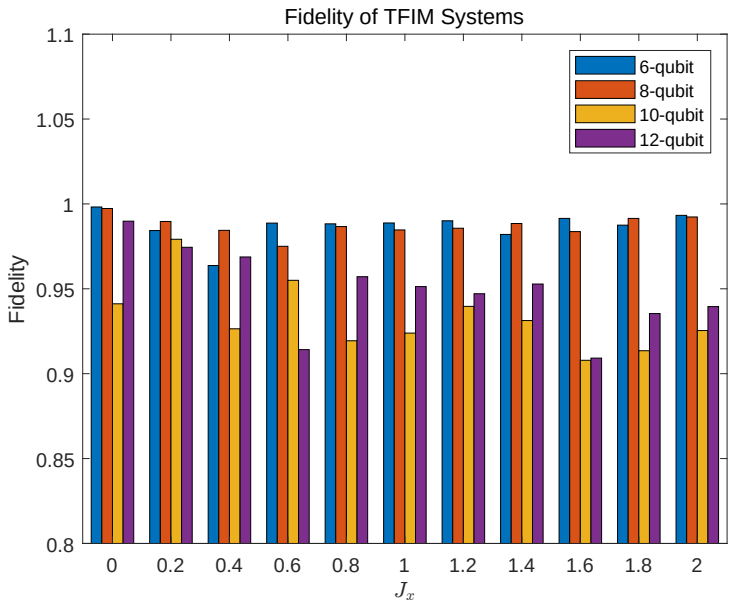

Figure 2. The prediction fidelity for different number of qubits. The fidelity is computed by averaging 400 random samples.

We test the neural network on quantum systems of 6 and 12 qubits at $J_x = 1$ using 600 random unitary matrices. The prediction fidelity under these transformations is shown as heatmaps in Figure 3. The model achieves prediction fidelity above 90%, validating its effectiveness in capturing the characteristics of the quantum state.

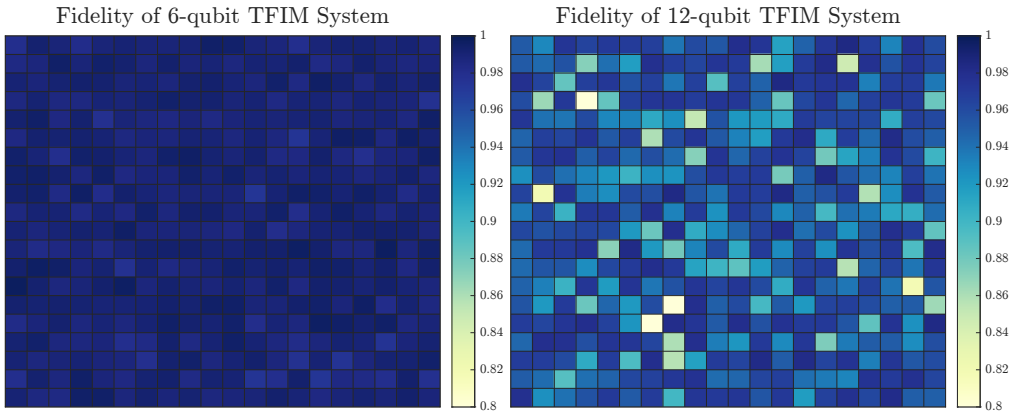

Figure 3. The performance of the neural network on TFIM states. (**a**) is for 6-qubit system, and (**b**) is for 12-qubit system. Both show the accuracy of the neural network predictions on randomly sampled unitary matrices.

We also examine the relationship between sample size and model performance at a fixed transverse field of 1.0. The model accuracy is evaluated using average fidelity across 400 random unitary matrices. Figure 4 indicates that high fidelity can be achieved with limited measurements: 6-qubit systems reach approximately 90% fidelity with 200 samples, while 12-qubit systems require approximately 600 samples for comparable performance. Overall, both

the model parameters and required training samples scale polynomially with qubit number, indicating efficient scaling of the neural network architecture.

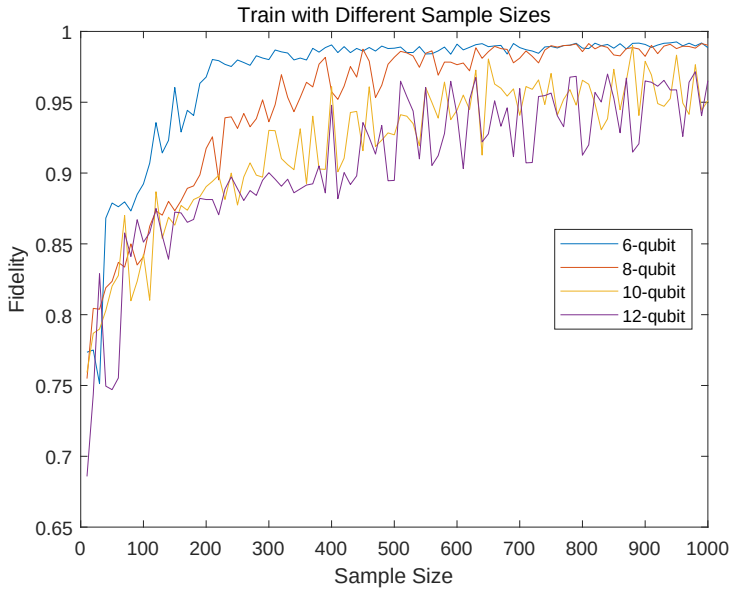

Figure 4. The average prediction accuracy as the number of samples increases. Generally, as the number of samples increases, the average prediction accuracy also improves.

Finally, we show how the minimum required sample size varies with qubit number for a given fidelity threshold, as in Figure 5. We use the same method as in Figure 4 to generate results for quantum systems ranging from 6 to 12 qubits, and define the first intersection point with the fidelity level line as the minimum sample size. We select two fidelity levels, 0.9 and 0.95, to plot. Due to the inherent randomness in the data generation and training process, the result is not strictly monotonically increasing as expected. However, the trend follows an approximately power-law relation with a power less than 2.

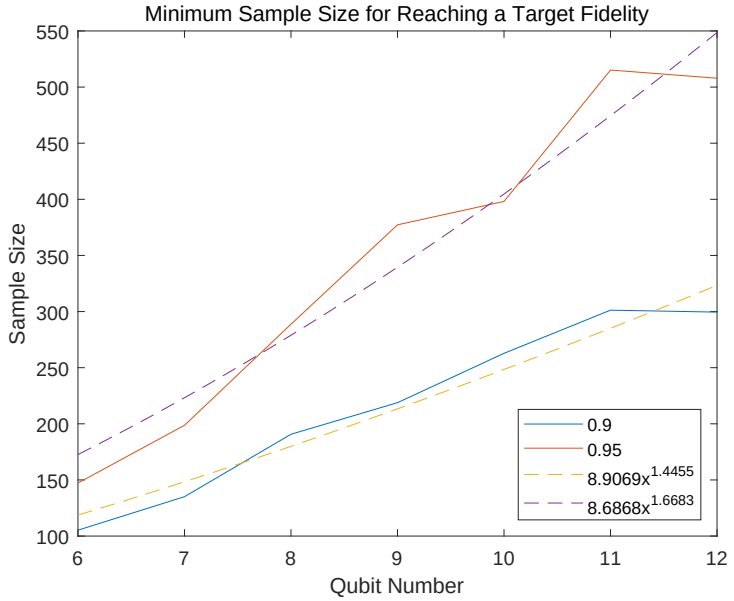

Figure 5. Relation between qubit number and sample size. The overall trend suggests a power-law relationship between the sample size and qubit number. The dashed lines represent the fitted curves for the corresponding data.

6.2. Comparison between Different Neural Network Models

Given that Cha et al. [25] showed the applicability of attention mechanisms to quantum state tomography, we conduct a comparative analysis between attention-based approaches and our proposed method. We investigate two

distinct attention-based implementations: (1) reformulating our quantum state reconstruction task as a sequence-to-sequence translation problem, and (2) substituting the LSTM architecture with a multi-head attention mechanism while maintaining the overall framework. To ensure fair comparison and prevent information leakage, we use causal masking in the attention mechanisms to maintain the sequential nature of quantum measurements. Our empirical results, as illustrated in Figure 6, demonstrate that our LSTM-based architecture outperforms both attention-based variants.

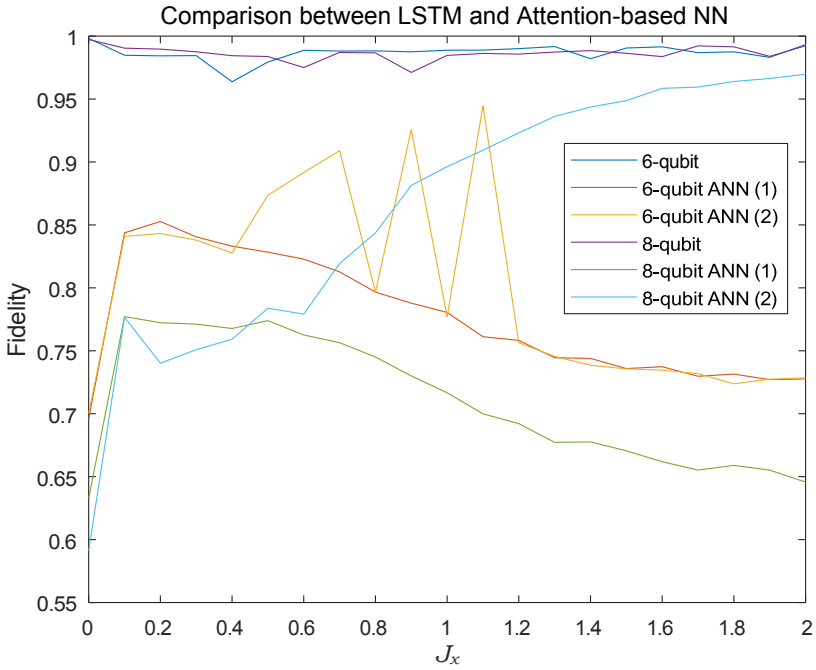

Figure 6. Performances of different neural network architecture. The fidelity is computed by averaging 400 random samples. It can be found that our LSTM model behaves well, but the attention-based neural networks (ANN (1) and ANN (2)) are unstable.

6.3. Other States

To further validate our findings, we extend our analysis to a more complex quantum system: the cat state from quantum optics, which has the form

$$|\phi\rangle_{cat} = |\alpha\rangle + |-\alpha\rangle,$$

where $|\alpha\rangle$ is the coherent state, and $\alpha \in \mathbb{C}$. While theoretically infinite-dimensional, this state can be effectively studied through dimensional truncation, providing an ideal test case for our method. We generate quantum cat states using the QuTiP library [26]. As illustrated in Figure 7, the neural network demonstrates robust performance in accurately representing this quantum state, further verifying the generalizability of our approach.

We also conduct experiments on two different quantum states. The first state is a randomly generated state using Qutip, with a dense density matrix. From the results in Figure 7, the neural network shows impressive performance in this state, with the fidelity of each measurement being very close to 1. The second state is a W-state with n^2 nonzero elements, where n is the qubit number. It has the following form:

$$|\phi\rangle_w = \frac{1}{\sqrt{n}}(|\underbrace{100\cdots0}_{n}\rangle + |010\cdots0\rangle + \cdots + |000\cdots1\rangle).$$

The results in this state show that the fidelity remains approximately above 0.9.

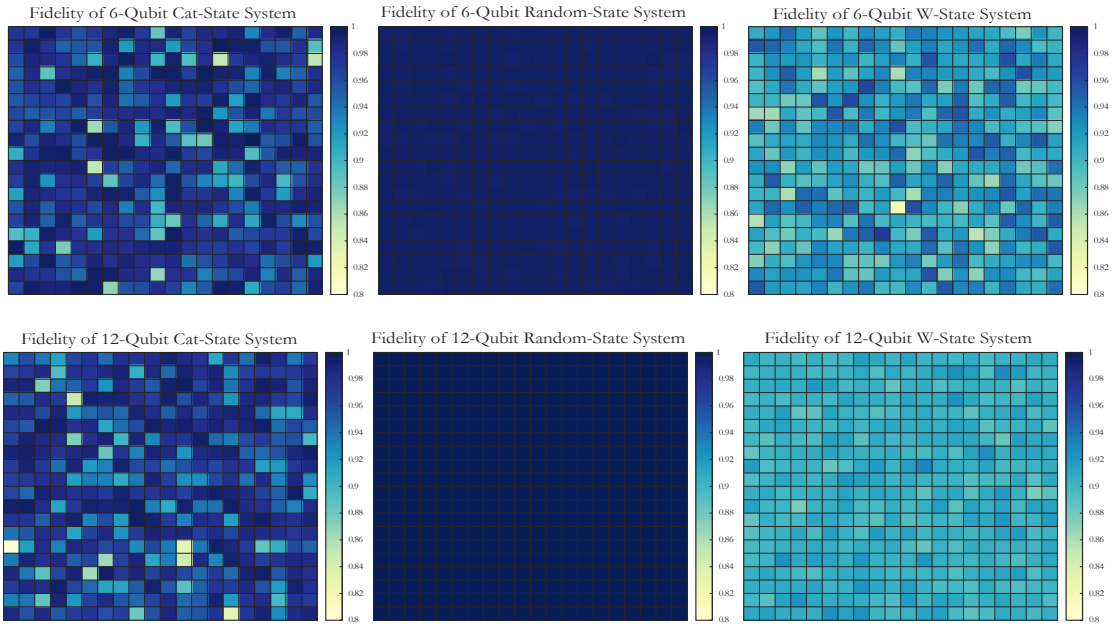

Figure 7. Performance of the neural network on different quantum states. Both of them displays the accuracy of the neural network predictions on randomly sampled unitary matrices. The first row consists of 6-qubit systems, while the second row comprises 12-qubit systems. From left to right, each column represents a different quantum state: cat-state, random state, and w-state.

7. Conclusion

This work presents a novel approach for quantum sample tomography that differentiates itself from conventional methods. We have developed both theoretical foundations and practical implementations for this challenging problem. Our analysis, grounded in local Rademacher complexity theory, establishes a fundamental theorem that rigorously justifies the application of machine learning method to quantum measurement prediction.

Building upon this mathematical foundation, we have developed a specialized Long Short-Term Memory (LSTM) architecture, incorporating symmetry-preserving constraints and invariance properties inherent to quantum systems. Our model maintains polynomial complexity with system size. Through several numerical experiments across various quantum systems, including the TFIM, we demonstrate that our approach consistently achieves good accuracy compared to other methods, including recent attention-based models. However, RNNs have several limitations. One challenge is the gradient propagation when processing long sequences of quantum states. Additionally, our model assumes noise-free conditions, and small perturbations in the quantum system may lead to huge changes in the RNN parameters. Future work is to improve the neural network performance for large-scale quantum systems.

Author Contributions

Kun Tang: Conceptualization, characterization, analysis, methodology, investigation, writing—original draft. **Jun Lai:** Conceptualization, methodology, supervision, writing—review and editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

The work of Jun Lai was supported by the National Natural Science Foundation of China (NSFC) under grant No. 12371427.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the work reported in this paper.

Data Availability Statement

The authors confirm that no external data were used in this study. All necessary data are fully available within the paper.

Acknowledgments

Jun Lai would like to acknowledge the “Xiaomi Young Scholars” program from Xiaomi Foundation.

References

1. F. Arute, K. Arya, R. Babbush, et al. (2019). "Quantum supremacy using a programmable superconducting processor". *Nature*, 574, 505–510.
2. A.W. Harrow, A. Hassidim, and S. Lloyd (2009). "Quantum algorithm for solving linear systems of equations." *Physical Review Letters*, 103, 150502.
3. L. Zhou, S.-T. Wang, S. Choi, et al. (2020). "Quantum approximate optimization algorithm: performance, mechanism, and implementation on near-term devices." *Physical Review X,* 10, 021067.
4. A.W.R. Smith, J. Gray, and M.S. Kim (2021). "Efficient quantum state sample tomography with basis-dependent neural networks." *PRX Quantum*, 2, 020348.
5. J. Shang, Z. Zhang, and H.K. Ng (2017). "Superfast maximum-likelihood reconstruction for quantum tomography." *Physical Review A*, 95, 062336.
6. E. Bolduc, G.C. Knee, E.M. Gauger, et al. (2017). "Projected gradient descent algorithms for quantum state tomography." *npj Quantum Information*, 3, 44.
7. D. Gross, Y.-K. Liu, S.T. Flammia, et al. (2010). "Quantum state tomography via compressed sensing." *Physical Review Letters*, 105, 150401.
8. Z. Qin, C. Jameson, Z. Gong, et al. (2024). "Quantum state tomography for matrix product density operators." *IEEE Transactions on Information Theory*, 70, 5030.
9. J. van Apeldoorn, A. Cornelissen, A. Gilyén, et al. (2023). "Quantum tomography using state-preparation unitaries", in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1265–1318.
10. G. Torlai, G. Mazzola, J. Carrasquilla, et al. (2018). "Neural-network quantum state tomography". *Nature Physics*, 14, 447–450.
11. S. Ahmed, C. Sánchez Muñoz, F. Nori, and A.F. Kockum (2021). "Classification and reconstruction of optical quantum states with deep neural networks". *Physical Review Research*, 3, 033278.
12. V. Wei, W.A. Coish, P. Ronagh, et al. (2024). "Neural-shadow quantum state tomography". *Physical Review Research*, 6, 023250.
13. Y. Quek, S. Fort, and H.K. Ng (2021). "Adaptive quantum state tomography with neural networks". *npj Quantum Information*., 7, 1–7.
14. N. Innan, O. I. Siddiqui, S. Arora, et al. (2024). "Quantum state tomography using quantum machine learning". *Quantum Machine Intelligence*, 6, 28.
15. S. Aaronson (2007). "The learnability of quantum states". *Proceedings of the Royal Society A*, 463, 3089–3114.
16. H.-Y. Hu, S. Choi, and Y.-Z. You (2023). "Classical shadow tomography with locally scrambled quantum dynamics". *Physical Review Research,* 5, 023027.
17. H.-Y. Huang (2022). "Learning quantum states from their classical shadows". *Nature Review Physics,* 4, 81.
18. P.L. Bartlett, O. Bousquet, and S. Mendelson (2005). "Local Rademacher complexities". *Annals of Statistics*, 33.
19. A. Kitaev, A. Shen, and M. Vyalyi (2002). *Classical and Quantum Computation*. American Mathematical Society.
20. J. Cotler and F. Wilczek (2020). "Quantum overlapping tomography". *Physical Review Letters*, 124, 100401.
21. S. Aaronson (2020). "Shadow tomography of quantum states". *SIAM Journal on Computing*, 49, STOC18-368-STOC18-394.
22. A. Maurer (2016). "A vector-contraction inequality for rademacher complexities", in R. Ortner, H.U. Simon, and S. Zilles (eds), *Algorithmic Learning Theory*. Cham: Springer International Publishing, 3–17.
23. C. Cortes, M. Kloft, and M. Mohri (2013). "Learning kernels using local Rademacher complexity". *Advances in Neural Information Processing Systems (NIPS)*, 26, 2760–2768.
24. S. Dasgupta and A. Gupta (2003). "An elementary proof of a theorem of Johnson and Lindenstrauss". *Random Structures & Algorithms*, 22, 60.
25. P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, et al. (2022). "Attention-based quantum tomography". *Machine Learning: Science and Technology*, 3, 01LT01.
26. J.R. Johansson, P.D. Nation, and F. Nori (2012). "QuTiP: An open-source python framework for the dynamics of open quantum systems". *Computer Physics Communications*, 183, 1760–1772.
27. J.R. Johansson, P.D. Nation, and F. Nori (2013). "QuTiP 2: A python framework for the dynamics of open quantum systems". *Computer Physics Communications*, 184, 1234–1240.

28. B.I. Bantysh, A.Y. Chernyavskiy, and Y.I. Bogdanov (2021). “Quantum tomography benchmarking”. *Quantum Information Processing*, 20, 339.
29. J. Carrasquilla, G. Torlai, R.G. Melko, et al. (2019). “Reconstructing quantum states with generative models”. *Nature Machine Intelligence*, 1, 155–161.