New Journal of Physics

The open access journal at the forefront of physics

CrossMark

PAPER

Full reconstruction of a 14-qubit state within four hours

Zhibo Hou[1,2], Han-Sen Zhong[1,2], Ye Tian[1,2], Daoyi Dong[3], Bo Qi[4], Li Li[5], Yuanlong Wang[3], Franco Nori[6,7], Guo-Yong Xiang[1,2], Chuan-Feng Li[1,2] and Guang-Can Guo[1,2]

1   Key Laboratory of Quantum Information, University of Science and Technology of China, CAS, Hefei, 230026, People's Republic of China
2   Synergetic Innovation Center of Quantum Information and Quantum Physics, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China
3   School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2600, Australia
4   Key Laboratory of Systems and Control, ISS, and National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, CAS, Beijing 100190, People's Republic of China
5   Centre for Quantum Computation and Communication Technology and Centre for Quantum Dynamics, Griffith University, Brisbane, Queensland 4111, Australia
6   CEMS, RIKEN, Saitama 351-0198, Japan
7   Physics Department, University of Michigan, Ann Arbor, Michigan 48109-1040, USA

E-mail: gyxiang@ustc.edu.cn

Abstract

Full quantum state tomography (FQST) plays a unique role in the estimation of the state of a quantum system without *a priori* knowledge or assumptions. Unfortunately, since FQST requires informationally (over)complete measurements, both the number of measurement bases and the computational complexity of data processing suffer an exponential growth with the size of the quantum system. A 14-qubit entangled state has already been experimentally prepared in an ion trap, and the data processing capability for FQST of a 14-qubit state seems to be far away from practical applications. In this paper, the computational capability of FQST is pushed forward to reconstruct a 14-qubit state with a run time of only 3.35 hours using the linear regression estimation (LRE) algorithm, even when informationally overcomplete Pauli measurements are employed. The computational complexity of the LRE algorithm is first reduced from $\sim 10^{19}$ to $\sim 10^{15}$ for a 14-qubit state, by dropping all the zero elements, and its computational efficiency is further sped up by fully exploiting the parallelism of the LRE algorithm with parallel Graphic Processing Unit (GPU) programming. Our result demonstrates the effectiveness of using parallel computation to speed up the postprocessing for FQST, and can play an important role in quantum information technologies with large quantum systems.

1. Introduction

Quantum state tomography [1–3], characterizing the state of a quantum system via quantum measurements and data processing, is a starting point and the standard for verification and benchmarking of various quantum information processing tasks, such as quantum computation [4], cryptography [5], and metrology [6–10].

To reconstruct quantum states wherein we have no *a priori* information, we can resort to informationally (over)complete measurements. Quantum state tomography [11, 12] using informationally (over)complete measurements is referred as *full quantum state tomography* (FQST) in this paper. As there are $(d^2 - 1)$ independent parameters to characterize the density matrix of a d-dimensional quantum state, FQST needs at least $(d^2 - 1)$ measurement operators. Note that the dimension d grows exponentially with the size of the quantum system. Thus, the number of measurements and the computational complexity of data processing in FQST suffer the curse of dimensionality. Moreover, the time of data processing was found to be even much longer than the time required for implementing the measurements. As reported in [13–15], reconstructing an 8-qubit state using maximum likelihood estimation (MLE) [1, 16] took almost a week, while the measurement

Figure 1. Reciprocal of the run time versus the number of employed streaming multiprocessors (SMs). Red dots denote the reciprocal of the run time in step (i) for a 12-qubit state when m SMs are used for parallel programming. Each SM has 192 CUDA cores. Since one thread is mapped to one CUDA core, m SMs have $192m$ threads running in parallel. The least-square linear fitting (red dotted line) shows that the speed of parallel programming increases almost proportionally with the number of parallel threads.

time was only 10 hours [13]. With the breakthrough and rapid development of experimental techniques, the size of quantum systems with entanglement or coherence prepared in the laboratory has already grown from 2 qubits [17, 18] to 10 qubits in photonic systems (e.g., 8 qubits in [19, 20] and 10 qubits in [21]), 12 qubits in NMR [22] and to even 14 qubits [23] in ion traps, overwhelming the capability of the available full quantum state tomography.

Significant effort has been devoted to improving the performance of quantum state tomography [24–28]. Some methods focused on extracting partial concerned information. For example, entanglement witness [29, 30] can detect entanglement with few measurements; direct purity estimation [31] and fidelity estimation [13, 32] were utilized to obtain the purity of the prepared state and its fidelity with the ideal state; permutationally invariant tomography [33] was used to extract information that will not change under permutation. Several approaches were concerned on performing quantum state tomography with *a priori* knowledge or assumptions. For example, compressed sensing [34–36] can perform quantum state tomography for quantum states with a low rank. If a quantum state is a matrix product state, it is possible to develop efficient tomography algorithms [37, 38]. However, these methods either extract partial information or have some prior information about the state to be reconstructed.

Several approaches have also been presented to reduce the computational complexity in the reconstruction algorithms in FQST [14, 39]. For example, the authors in [14] developed an algorithm which can be used to efficiently reconstruct a 9-qubit state in about five minutes. However, when the size of the quantum system increases one qubit, the running time will increase by a factor of more than ten according to figure 1 in [14], resulting in years of computation time for a 14-qubit state. In [39], a linear regression estimation (LRE) algorithm was proposed which has a much lower computational complexity than that of MLE for quantum state tomography [40]. In this paper, we push the data processing capability of FQST to a 14-qubit state using the informationally overcomplete Pauli measurements by optimizing the LRE algorithm in [39] and employing parallel programming with graphics processing unit (GPU).

For experimental ease and high level of estimation accuracy, Pauli measurements are the preferred choice in experiments of FQST, although they are informationally overcomplete. LRE was demonstrated to be much more efficient than MLE in FQST [39]. However, in order to reconstruct a 14-qubit state efficiently, we need to further optimize the LRE method. The efficiency of FQST in this paper refers solely to the reduced computational complexity of data processing. Our first optimization is based on the fact that the representation of Pauli bases in the algorithm has very few nonzero elements under a proper choice of the representation. Furthermore, thanks to the simple LRE algorithm which only involves additions and multiplications of vectors and matrices, it is naturally suitable to be sped up by parallel programming. In parallel programming of matrix computation, GPU works much better than the central processing unit (CPU). Hence, we can use GPU parallel programming to realize the LRE algorithm and enhance the FQST capability. Compared with the result in [39], in this paper we optimize the LRE method with reduced computational complexity and storage requirement, and implement the optimized LRE method using GPU parallel programming. The results show significant enhancement of the FQST capability.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to the three steps of LRE and analyzes the computational complexity and the storage requirement in the case of informationally complete measurements. In section 3, computational complexity and storage requirement are discussed based on Pauli measurements. In section 4, the LRE algorithm in the first two steps is realized using parallel GPU programming. The run time of the algorithm with GPU speeding up is also compared with that using CPU programming. In

section 5, the estimation error of reconstructing a maximally-mixed state is analyzed in terms of squared Hilbert-Schmidt (HS) distance and infidelity. Section 6 presents the summary and prospect of this paper.

2. Linear regression estimation algorithm

In linear regression estimation (LRE) for a d-dimensional quantum system, the density matrix and measurement bases take a vector form after we choose a representation basis set $\{\Omega_i\}_{i=0}^{d^2-1}$ [39]. The operators in this basis set are orthonormal, i.e., $\mathrm{Tr}(\Omega_i^\dagger \Omega_j) = \delta_{ij}$. For convenience, let $\Omega_i = \Omega_i^\dagger$ and all the bases are traceless except $\Omega_0 = (1/d)^{\frac{1}{2}} I$. Elements in the vector form Θ of a quantum state ρ are given by

$$\theta_i = \mathrm{Tr}(\rho \Omega_i). \tag{1}$$

Given a set of M measurement operators $\{|\Psi\rangle\langle\Psi|^{(j)}\}_{j=1}^M$, elements in the vector form $\Gamma^{(j)}$ of each $|\Psi\rangle\langle\Psi|^{(j)}$ are given by

$$\gamma_i^{(j)} = \mathrm{Tr}(|\Psi\rangle\langle\Psi|^{(j)} \Omega_i). \tag{2}$$

The whole LRE algorithm consists of three steps: step (i) Obtain the estimate of Θ using measurement data; step (ii) Construct a Hermitian matrix $\hat{\mu}$ satisfying $\mathrm{Tr}\,\hat{\mu} = 1$ from the estimate of Θ; and step (iii) Find a physical density matrix $\hat{\rho}$ close to $\hat{\mu}$.

In step (i), a least-squared estimate $\hat{\Theta}^{LS}$, without consideration of the positivity restriction of quantum states, is given by

$$\hat{\Theta}^{LS} = (X^\top X)^{-1} \sum_{j=1}^M \hat{p}_j \Gamma^{(j)}, \tag{3}$$

where $\hat{p}_j$ is the measured frequency of $|\Psi\rangle\langle\Psi|^{(j)}$, $X = \left(\Gamma^{(1)}, \cdots, \Gamma^{(M)}\right)^\top$ and $X^\top X = \sum_{j=1}^M \Gamma^{(j)} \Gamma^{(j)\top}$. The computational complexity in this step is at least $O(d^4)$, since FQST requires an informationally complete measurement set $\{|\Psi\rangle\langle\Psi|^{(j)}\}_{j=1}^M$ with $M = d^2 - 1$. For a 14-qubit state, $d = 2^{14}$ and the computational complexity is $\sim 10^{16}$.

In step (ii), on the basis of the solution $\hat{\Theta}^{LS}$ to (3) in step (i), we can obtain a Hermitian matrix $\hat{\mu}$ with $\mathrm{Tr}\,\hat{\mu} = 1$ by

$$\hat{\mu} = \sum_{i=0}^{d^2-1} \hat{\theta}_i^{LS} \Omega_i. \tag{4}$$

The computational complexity in this step is also $O(d^4)$.

The state estimate $\hat{\mu}$ obtained in (4) may have negative eigenvalues and be nonphysical due to the randomness of the finite measurement results. In step (iii), a proper method needs to be adopted to pull $\hat{\mu}$ back to a physical state. In this step, we use the fast algorithm proposed in [14], where a physical estimate $\hat{\rho}$ is chosen to be the closest density matrix to $\hat{\mu}$ under the matrix 2-norm. According to [14], the computational complexity in this step is $O(d^3)$.

It is clear that the computational complexity in LRE is dominated by the first two steps, which is $O(d^4)$. For a 14-qubit state, this is $\sim 10^{16}$. In terms of storage, $O(d^4)$ and $O(d^2)$ bytes are required to store all the measurement bases and measurement results, respectively, for an informationally complete measurement set. For a 14-qubit state, this needs tens of thousands of terabytes, which is beyond the capability for practical applications. In the following, we develop a method to reduce the computational complexity and storage requirement.

3. Computational complexity and storage with Pauli measurements

Pauli measurements are a good choice to extract information in quantum state tomography of n-qubit systems ($d = 2^n$) because of not only their experimental ease but also the ability to achieve a high level of accuracy. With all the possible combinations of Pauli measurements for n-qubit systems, the total number of measurement bases is $M = 6^n$. Without optimization, this informationally overcomplete measurement set further increases the computational complexity in step (i) to $\sim 10^{19}$ for a 14-qubit state and also increases the storage requirement from $\sim 10^{16}$ to $\sim 10^{19}$.

The computational complexity and storage requirement can be greatly reduced because many terms in $\Gamma^{(j)}$ are zero when $\{\Omega_i\}_{i=0}^{d^2-1}$ are chosen as the tensor product of $\{\frac{1}{\sqrt{2}}\sigma_i\}_{i=0}^3$, with $\sigma_0 = I_{2\times 2}$, and Pauli matrices

$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$, and $\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. For an n-qubit state,

$$\Omega_i = 2^{-n/2} \bigotimes_{k=1}^{n} \sigma_{i_k} \tag{5}$$

with $i_k = 0, 1, 2, 3$, and $i = \sum_{k=1}^{n} i_k \times 4^{n-k}$.

In the 1-qubit scenario, inserting (5) into (2), one can easily obtain $\Gamma^{(j)}$ for all the eigenvectors of the three Pauli operators σ_1, σ_2, and σ_3, which are $\frac{1}{\sqrt{2}}(1, \pm 1, 0, 0)^T$, $\frac{1}{\sqrt{2}}(1, 0, \pm 1, 0)^T$, and $\frac{1}{\sqrt{2}}(1, 0, 0, \pm 1)^T$. The term $X^\top X$ is calculated to be a diagonal matrix with $X^\top X = \mathrm{diag}\{3, 1, 1, 1\}$. As for the n-qubit scenario, since the measurement bases are the tensor product of those for single qubits, $\Gamma^{(j)}$ and $X^\top X$ are also the tensor products of their counterparts for relevant 1-qubit cases. Thus, there will be only 2^n nonzero elements in $\Gamma^{(j)}$, instead of the original 4^n, and $X^\top X$ are diagonal. Dropping all zero elements will reduce the computational complexity in step (i) from $O(24^n)$ to $O(12^n)$, i.e., from $\sim 10^{19}$ to $\sim 10^{15}$, for a 14-qubit state.

After the optimization in step (i), the computational complexity $O(16^n)$ in step (ii) is dominant. Recall the definition of Ω_i in (5), there are only 2^n nonzero elements. Therefore, after dropping all zero elements, the computational complexity in step (ii) can be reduced to $O(8^n)$ (i.e., $\sim 10^{12}$) for a 14-qubit state. Since the computational complexity in step (iii) is only $O(8^n)$, the total computational complexity with Pauli measurements is $O(12^n)$.

In terms of storage, without dropping the zero elements, $O(24^n)$ bytes are needed to store all the measurement bases. Even after dropping all the zero elements, the storage requirement is still $O(12^n)$ for the nonzero elements and their locations. However, the storage requirement of the measurement bases can be further reduced to $O(6^n)$ after the 6^n bases are divided into 3^n groups. The 2^n operators in each group are all the eigenvectors of one combination of three Pauli operators. The 2^n nonzero elements in each of these 2^n operators in the same group share the same locations. All the nonzero values of operators in one group form a $2^n \times 2^n$ nonzero matrix, which is the same for all the 3^n groups. Hence, we only need to store 3^n types of locations of nonzero elements (with $O(6^n)$ storage requirement), and a $2^n \times 2^n$ nonzero matrix (with $O(4^n)$ storage requirement). Besides, the storage requirement for the measurement results of all the 6^n bases also requires $O(6^n)$. Thus, all the needed storage is only $O(6^n)$. Thus, the storage cost in the 14-qubit scenario is $\sim 10^{10}$, which is only tens of Giga Bytes.

4. Parallel GPU programming

Thanks to the direct and simple formula of the LRE algorithm in (3) and (4), only addition and multiplication operations of matrices or vectors are involved, which are naturally suitable for parallel programming. Graphic processing units possess powerful capability for parallel programming. This technique is exploited to speed up the computation in both step (i) and step (ii). The computer hardware includes 500 GB hard drive, 16 GB memory, i7-4770 CPU with 3.5 GHz, 4 cores and 8 threads, and GTX780 GPU with 2304 CUDA cores and 3G standard memory.

As analyzed in section 3, step (i) has the dominant computational complexity $O(12^n)$. GPU parallel programming is employed in this step. According to the locations of nonzero elements, all the measurement bases can be divided into 3^n groups. Each group has 2^n measurement bases. In parallel GPU programming, firstly, one group of measurement bases and their measurement data are put in, wherein the data include a $2^n \times 2^n$ nonzero matrix, 2^n nonzero locations and 2^n measurement results. To be parallel, each thread is devoted to calculating one element of $\hat{\Theta}^{LS}$, whose location is in the 2^n nonzero locations. Hence, only 2^n elements of the total 4^n elements in $\hat{\Theta}^{LS}$ need to be calculated and updated for one group of data. All the threads are synchronized after updating these 2^n elements of $\hat{\Theta}^{LS}$ to get ready for the computation of the next group of data. Then this process continues until all the 3^n groups of data are computed.

In order to show the relationship between the amount of speed up and the number of parallel threads in GPU programming, the reciprocal of the run time in step (i) for a maximally-mixed 12-qubit state is plotted with respect to m in figure 1, where m is the number of streaming multiprocessors (SMs). Here, we only consider step (i) for simplicity, because step (i) has the dominant computational complexity. There are 12 SMs in the GTX780 GPU and each SM contains 192 CUDA cores. Since each CUDA core is occupied by a thread, m SMs can execute $192m$ threads in parallel. Figure 1 shows that the speed in step (i), which is denoted by the reciprocal of the time cost, increases almost proportionally with the number of parallel threads, represented by the number of SMs. Hence, all the SMs are used in GPU programming to gain the fastest speed in the rest simulations.

Using this approach we now demonstrate the computation time of reconstructing multi-qubit states using our algorithm. The computation time of GPU programming in step (i) costs 2.78 hours for a 14-qubit state as depicted in figure 2(a). In comparison, to reconstruct a 11-qubit state, the CPU programming has taken almost half an hour for step (i) and will be difficult to compute larger systems as shown in figure 2(a).

Figure 2. Computation time with respect to the number of qubits. Subfigures (a) to (d) depict, respectively, the computation time in steps (i)–(iii) and the whole process in the numerical reconstruction of a maximally-mixed state. Blue crosses and red dots represent the computation time for CPU and GPU programming of our optimized algorithm, respectively. When the number of qubits is larger than 8, GPU programming begins to show its advantage over CPU programming. The dashed blue line is an exponential fitting of all the blue crosses. The dotted red line is an exponential fitting of the last four red dots because the time cost of GPU programming for a small sized system is dominated by the overhead of kernel calls.

In step (ii), the time cost of CPU programming for a 14-qubit state approximates to be 10 hours according to the numerical fitting line (blue dashed line) in figure 2(b), which is much longer than the run time of step (i) via GPU programming. Hence, GPU programming is also used to speed up the computation in this step. We note that the nonzero elements in $I_{2\times2}$ and σ_z have the same locations, and so it is for σ_x and σ_y. Thus all the d^4 Ω_i, which are constructed in (5) by the above four matrices, can be divided into 2^n groups according to the same locations of nonzero elements. In parallel GPU programming, one group of 2^n Ω_i and the corresponding 2^n elements in $\hat{\Theta}^{LS}$ are put in. Each thread is responsible to calculate one of the corresponding 2^n elements in $\hat{\mu}$. All the threads are synchronized after finishing the update of the 2^n elements in $\hat{\mu}$ to get prepared for the next group of data. Then this step is repeated until all the 2^n groups of data are calculated. With parallel GPU programming, the computation time in step (ii) is reduced to only 0.08 hours, as shown in figure 2(b).

In step (iii), we do not need GPU programming since the time cost of CPU programming with Mathematica is only 0.49 hours (see figure 2(c)) for a 14-qubit state, which is already about 5 times shorter than the computation time in step (i) using GPU programming. The total computation time of the whole process in reconstructing a 14-qubit state turns out to be only 3.35 hours, as shown in figure 2(d). From figure 2(d), we know that our optimized LRE algorithm based on CPU programming (blue) is already more than 100 times faster than the efficient algorithm in [14] for reconstructing a 9-qubit state. However, according to the numerical fitting line in figure 2(d), reconstructing a 14-qubit state will take more than a month using our optimized algorithm with CPU programming. This is also a reason why we resort to GPU programming for further speeding up through considering the parallelism of our algorithm. In our numerical simulations, all states are chosen as the maximally-mixed states in n-qubit systems only for the ease of obtaining the simulated measurement frequency. Our method is applicable to other states because the computational complexity and run time are state independent.
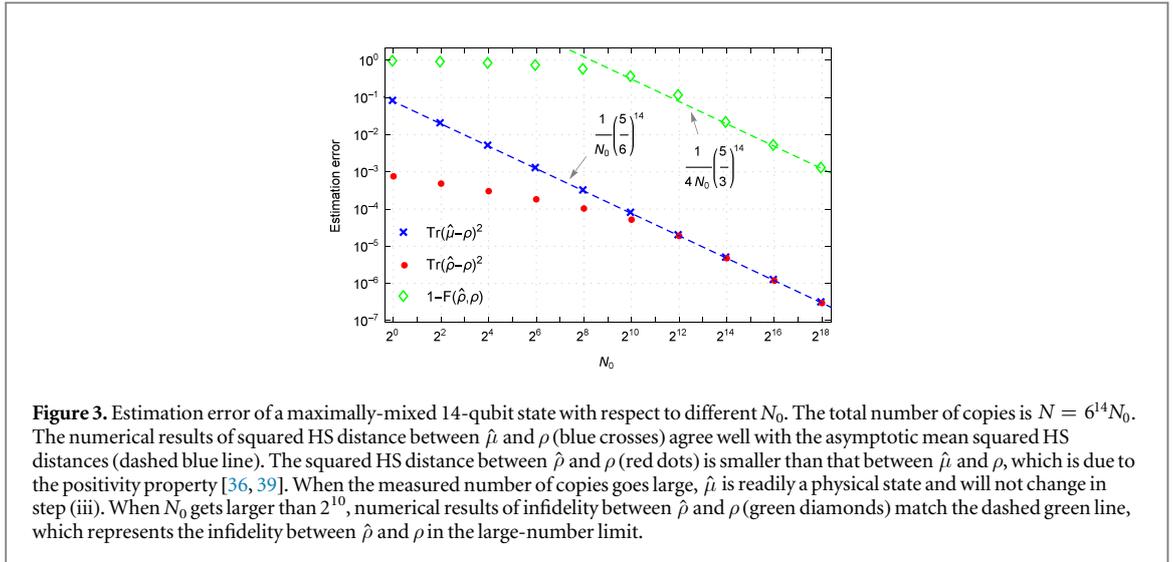
Figure 3. Estimation error of a maximally-mixed 14-qubit state with respect to different N_0. The total number of copies is $N = 6^{14}N_0$. The numerical results of squared HS distance between $\hat{\mu}$ and ρ (blue crosses) agree well with the asymptotic mean squared HS distances (dashed blue line). The squared HS distance between $\hat{\rho}$ and ρ (red dots) is smaller than that between $\hat{\mu}$ and ρ, which is due to the positivity property [36, 39]. When the measured number of copies goes large, $\hat{\mu}$ is readily a physical state and will not change in step (iii). When N_0 gets larger than 2^{10}, numerical results of infidelity between $\hat{\rho}$ and ρ (green diamonds) match the dashed green line, which represents the infidelity between $\hat{\rho}$ and ρ in the large-number limit.

5. Estimation error

In this section, the error between the estimate and the real state is discussed in terms of the squared Hilbert-Schmidt (HS) distance and infidelity. The mean squared HS distance between the estimate state $\hat{\mu}$ in step (ii) and the true state ρ is $\mathrm{E}\mathrm{Tr}(\hat{\mu} - \rho)^2 = \mathrm{E}(\hat{\Theta}^{LS} - \Theta)^\top (\hat{\Theta}^{LS} - \Theta)$, asymptotically given by

$$\mathrm{E}\mathrm{Tr}(\hat{\mu} - \rho)^2 = \frac{M}{Nd}\mathrm{Tr}[(X^\top X)^{-1}X^\top PX(X^\top X)^{-1}], \tag{6}$$

where N is the total number of copies for all the M operators and $P = \mathrm{diag}(p_1 - p_1^2, \cdots, p_M - p_M^2)$, with $p_j = \mathrm{Tr}(\rho|\Psi\rangle\langle\Psi|^{(j)})$. Here we have one more factor of $1/d$ than equation (9) in [39] since every d measurement bases form a complete set of POVM and can be measured simultaneously. As these d measurement bases are measured simultaneously, the corresponding off-diagonal elements in P are of $O(p_ip_j)$ rather than zero. But these off-diagonal elements are much smaller than the diagonal elements when the size of the quantum system is large. Therefore, P is still approximated to only have diagonal elements in the case of simultaneous measurements.

This mean squared HS distance obviously depends on the unknown state. Here we calculate the mean squared HS distance when ρ is chosen as the maximally-mixed state, i.e., $\rho = \frac{I_{d\times d}}{d}$. One reason for this state choice is that an elegant theoretical result can be derived due to the simple form of a maximally-mixed state; the other reason is that the maximally-mixed state often gives the larger mean squared HS distance than the other states. In this case, we have $P = \frac{I_{d\times d}}{d}$ in the first order of approximation. When Pauli measurements are performed on copies of a n-qubit maximally-mixed state, the mean squared HS distance (blue line in figure 3) is equal to $\frac{1}{N_0}(\frac{5}{6})^n$ with $N_0 = \frac{N}{M}$. This theoretical result agrees well with the numerical results of $\mathrm{Tr}(\hat{\mu} - \rho)^2$ (blue crosses in figure 3). Thanks to the positivity condition, compared with $\mathrm{Tr}(\hat{\mu} - \rho)^2$, the squared HS distance of a physical estimate $\hat{\rho}$ for the real state (denoted as red dots in figure 3) is further reduced for $N_0 \leqslant 2^{10}$ although it remains the same when $N_0 > 2^{10}$ for a 14-qubit maximally-mixed state.

Another well-motivated figure of merit is infidelity, defined as $1 - F(\hat{\rho}, \rho) = 1 - \mathrm{Tr}^2(\sqrt{\sqrt{\rho}\hat{\rho}\sqrt{\rho}})$. When ρ is chosen as the maximally-mixed state, infidelity is directly related to the squared HS distance by $1 - F(\hat{\rho}, \rho) = \frac{d}{4}\mathrm{Tr}(\hat{\rho} - \rho)^2 + d^2 O(\mathrm{Tr}(\hat{\rho} - \rho)^3)$. In the large-number limit of N_0, $\hat{\mu}$ has non-negative eigenvalues and we have $\hat{\mu} = \hat{\rho}$. Thus, the average infidelity is well approximated by $\frac{1}{4N_0}(\frac{5}{3})^n$ (green line) in this limit. However, when N_0 is small, $\hat{\mu}$ will have negative eigenvalues and needs to be pulled back to a physical state $\hat{\rho}$. Hence, as shown in figure 3, the real infidelity (green diamonds) is much smaller than the prediction (green line) for $N < 2^{10}$ in the 14-qubit scenario.

6. Summary and prospect

In this paper, we have fully reconstructed a 14-qubit state with a modest computation time of 3.35 hours using informationally overcomplete Pauli measurements. By a smart choice of the representation basis set in the algorithm, only very few nonzero elements exist and need to be stored and calculated, which reduce the computational complexity by a factor of 2^n. Furthermore, parallel GPU programming is used to fully exploit the

parallelism in the simple LRE algorithm. It is worth pointing out that our method can push the capability of FQST forward to even larger quantum systems. This is because the measurement bases in our analysis and simulations are chosen as Pauli measurements, which are overcomplete. If we reduce the number of measurement bases from 6^n to 4^n, the computational complexity can be reduced to $O(8^n)$ from $O(12^n)$, which can further enhance the capability of FQST via our optimized LRE.

Acknowledgments

References

[1] Paris M G A and Řeháček J 2004 *Quantum State Estimation* (*Lecture Notes in Physics* vol 649) (Berlin: Springer)
[2] Hayashi M 2005 *Asymptotic Theory of Quantum Statistical Inference* (Singapore: World Scientific)
[3] Lvovsky A I and Raymer M G 2009 Continuous-variable optical quantum-state tomography *Rev. Mod. Phys.* **81** 299
[4] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[5] Gisin N, Ribordy G, Tittel W and Zbinden H 2002 Quantum cryptography *Rev. Mod. Phys.* **74** 145
[6] Giovannetti V, Lloyd S and Maccone L 2004 Quantum-enhanced measurements: beating the standard quantum limit *Science* **306** 1330
[7] Giovannetti V, Lloyd S and Maccone L 2006 Quantum metrology *Phys. Rev. Lett.* **96** 010401
[8] Giovannetti V, Lloyd S and Maccone L 2011 Advances in quantum metrology *Nat. Photon.* **5** 222
[9] Xiang G Y, Higgins B L, Berry D W, Wiseman H M and Pryde G J 2011 Entanglement-enhanced measurement of a completely unknown optical phase *Nat. Photonics* **5** 43
[10] Crowley P J D, Datta A, Barbieri M and Walmsley I A 2014 Tradeoff in simultaneous quantum-limited phase and loss estimation in interferometry *Phys. Rev. A* **89** 023845
[11] Liu Y X, Wei L F and Nori F 2004 Quantum tomography for solid-state qubits *Europhys. Lett.* **67** 874
[12] Liu Y X, Wei L F and Nori F 2005 Tomographic measurements on superconducting qubit states *Phys. Rev. B* **72** 014547
[13] da Silva M P, Landon-Cardinal O and Poulin D 2011 Practical characterization of quantum devices without tomography *Phys. Rev. Lett.* **107** 210404
[14] Smolin J A, Gambetta J M and Smith G 2012 Efficient method for computing the maximum-likelihood quantum state from measurements with additive gaussian noise *Phys. Rev. Lett.* **108** 070502
[15] Häffner H *et al* 2005 Scalable multiparticle entanglement of trapped ions *Nature* **438** 643
[16] James D F V, Kwiat P G, Munro W J and White A G 2001 Measurement of qubits *Phys. Rev. A* **64** 052312
[17] Kwiat P G, Mattle K, Weinfurter H, Zeilinger A, Sergienko A V and Shih Y 1995 New high-intensity source of polarization-entangled photon pairs *Phys. Rev. Lett.* **75** 4337
[18] Kwiat P G, Waks E, White A G, Appelbaum I and Eberhard P H 1999 Ultrabright source of polarization-entangled photons *Phys. Rev. A* **60** R773
[19] Huang Y F, Liu B H, Peng L, Li Y H, Li L, Li C F and Guo G C 2011 Experimental generation of an eight-photon greenberger-horne-zeilinger state *Nat. Commun.* **2** 546
[20] Yao X C, Wang T X, Xu P, Lu H, Pan G S, Bao X H, Peng C Z, Lu C Y, Chen Y A and Pan J W 2012 Observation of eight-photon entanglement *Nat. Photon.* **6** 225
[21] Gao W B, Lu C Y, Yao X C, Xu P, Gühne O, Goebel A, Chen Y A, Peng C Z, Chen Z B and Pan J W 2010 Experimental demonstration of a hyper-entangled ten-qubit schrödinger cat state *Nat. Phys.* **6** 331
[22] Negrevergne C, Mahesh T S, Ryan C A, Ditty M, Cyr-Racine F, Power W, Boulant N, Havel T, Cory D G and Laflamme R 2006 Benchmarking quantum control methods on a 12-qubit system *Phys. Rev. Lett.* **96** 170501
[23] Monz T, Schindler P, Barreiro J T, Chwalla M, Nigg D, Coish W A, Harlander M, Hänsel W, Hennrich M and Blatt R 2011 14-qubit entanglement: Creation and coherence *Phys. Rev. Lett.* **106** 130506
[24] Wang X B, Yu Z W, Hu J Z, Miranowicz A and Nori F 2013 Efficient tomography of quantum-optical gaussian processes probed with a few coherent states *Phys. Rev. A* **88** 022101
[25] Miranowicz A, Bartkiewicz K, Perina-Jr J, Koashi M, Imoto N and Nori F 2014 Optimal two-qubit tomography based on local and global measurements: Maximal robustness against errors as described by condition numbers *Phys. Rev. A* **90** 062123
[26] Miranowicz A, Özdemir S K, Bajer J, Yusa G, Imoto N, Hirayama Y and Nori F 2015 Quantum state tomography of large nuclear spins in a semiconductor quantum well: Optimal robustness against errors as quantified by condition numbers *Phys. Rev. B* **92** 075312
[27] Bartkiewicz K, Černoch A, Lemr K and Miranowicz A 2016 Priority choice experimental two-qubit tomography: measuring one by on all elements of density matrices *Sci. Rep.* **6** 19610
[28] Hou Z, Zhu H, Xiang G Y, Li C F and Guo G C 2016 Achieving quantum precision limit in adaptive qubit state tomography *npj Quantum Information* **2** 16001
[29] Barbieri M, De Martini F, Di Nepi G, Mataloni P, D'Ariano G M and Macchiavello C 2003 Detection of entanglement with polarized photons: Experimental realization of an entanglement witness *Phys. Rev. Lett.* **91** 227901
[30] Bourennane M, Eibl M, Kurtsiefer C, Gaertner S, Weinfurter H, Gühne O, Hyllus P, Bruß D, Lewenstein M and Sanpera A 2004 Experimental detection of multipartite entanglement using witness operators *Phys. Rev. Lett.* **92** 087902
[31] Bagan E, Ballester M A, Muñoz Tapia R and Romero-Isart O 2005 Purity estimation with separable measurements *Phys. Rev. Lett.* **95** 110504
[32] Flammia S T and Liu Y K 2011 Direct fidelity estimation from few Pauli measurements *Phys. Rev. Lett.* **106** 230501

[33] Tóth G, Wieczorek W, Gross D, Krischek R, Schwemmer C and Weinfurter H 2010 Permutationally invariant quantum tomography *Phys. Rev. Lett.* **105** 250403

[34] Gross D, Liu Y K, Flammia S T, Becker S and Eisert J 2010 Quantum state tomography via compressed sensing *Phys. Rev. Lett.* **105** 150401

[35] Liu W T, Zhang T, Liu J Y, Chen P X and Yuan J M 2012 Experimental quantum state tomography via compressed sampling *Phys. Rev. Lett.* **108** 170403

[36] Kalev A, Kosut R L and Deutsch I H 2015 Quantum tomography protocols with positivity are compressed sensing protocols *npj Quantum Information* **1** 15018

[37] Baumgratz T, Gross D, Cramer M and Plenio M B 2013 Scalable reconstruction of density matrices *Phys. Rev. Lett.* **111** 020401

[38] Cramer M, Plenio M B, Flammia S T, Somma R, Gross D, Bartlett S D, Landon-Cardinal O, Poulin D and Liu Y K 2010 Efficient quantum state tomography *Nat. Commun.* **1** 149

[39] Qi B, Hou Z, Li L, Dong D, Xiang G and Guo G 2013 Quantum state tomography via linear regression estimation *Sci. Rep.* **3** 3496

[40] Hou Z, Xiang G, Dong D, Li C F and Guo G C 2015 Realization of mutually unbiased bases for a qubit with only one wave plate: theory and experiment *Opt. Express* **23** 10018