Local Stochastic Factored Gradient Descent for Distributed Quantum State Tomography

Junhyung Lyle Kim, *Member, IEEE*, Mohammad Taha Toghani, *Member, IEEE*,
César A. Uribe, *Member, IEEE*, and Anastasios Kyrillidis, *Member, IEEE*

Abstract—We propose a distributed Quantum State Tomography (QST) protocol, named Local Stochastic Factored Gradient Descent (Local SFGD), to learn the low-rank factor of a density matrix over a set of local machines. QST is the canonical procedure to characterize the state of a quantum system, which we formulate as a stochastic non-convex smooth optimization problem. Physically, the estimation of a low-rank density matrix helps characterizing the amount of noise introduced by quantum computation. Theoretically, we prove the local convergence of Local SFGD for a general class of restricted strongly convex/smooth loss functions. Local SFGD converges locally to a small neighborhood of the global optimum at a linear rate with a constant step size, while it locally converges exactly at a sub-linear rate with diminishing step sizes. With a proper initialization, local convergence results imply global convergence. We validate our theoretical findings with numerical simulations of QST on the Greenberger-Horne-Zeilinger (GHZ) state.

Index Terms—Distributed optimization, matrix factorization, non-convex optimization, quantum state tomography.

I. INTRODUCTION

A FULLY-FUNCTIONAL fault-tolerant quantum computer faces many technical hurdles. For instance, using superconducting materials technology, quantum computers must remain cooled at a very low temperature—almost absolute zero—to preserve coherence [1]. Moreover, environmental noise from the electronics controlling the quantum system can disrupt the coherence of its qubits. Thus, the behavior of current quantum computer implementations needs to be characterized, verified, and certified, before their widespread commercial use [2].

Quantum State Tomography (QST) is the canonical procedure to characterize the state of a quantum system at various steps of a given computation [3]. In particular, experimental quantum physicists design quantum circuits that in theory lead to a specific target pure state; then, they compare the prepared (input) state and the reconstructed (output) state. To do so, measurements are taken on independently prepared copies of the state of a quantum system, and then used to estimate the unknown state by post-processing the data [3]. However, the description complexity of a quantum state grows *exponentially with the number of qubits*, leading to challenging data acquisition, processing, and storage. Therefore, as the number of qubits and quantum gates increases, so does the need for efficient, robust, and experimentally-accessible protocols to benchmark quantum information processors.

A quantum state can be represented by a density matrix ρ which is a complex, positive semi-definite (PSD) matrix with unit trace. The goal of QST is to design protocols that estimate ρ. For an n-qubit mixed state[1] $\Psi \in \mathbb{C}^{2^n}$, its density matrix can be written as a mixture of r pure states: $\rho = \sum_k^r p_k \Psi_k \Psi_k^\dagger \in \mathbb{C}^{2^n \times 2^n}$, where $(\cdot)^\dagger$ denotes the complex conjugate operator. Here, p_k is the probability of finding ρ in the pure state Ψ_k. Given these definitions, QST can be formulated as the estimation of a low-rank density matrix $\rho^\star \in \mathbb{C}^{d \times d}$ on an n-qubit Hilbert space with dimension $d = 2^n$, through the following ℓ_2-norm optimization problem:

$$\min_{\rho \in \mathbb{C}^{d \times d}} \quad F(\rho) := \frac{1}{2m} \|\mathcal{A}(\rho) - y\|_2^2$$
$$\text{subject to} \quad \rho \succeq 0, \ \text{rank}(\rho) \leq r, \tag{1}$$

where $\mathcal{A} : \mathbb{C}^{2^n \times 2^n} \to \mathbb{R}^m$ is the linear sensing map such that $\mathcal{A}(\rho)_k = \text{Tr}(A_k \rho)$, for $k = 1, \ldots, m$. The sensing map used in QST has a particular structure: it is the Kronecker product of Pauli matrices A_k, and is closely related to how quantum computers take measurements in practice [4].

The exponential dependency on the number of qubits implies that ρ has more than a trillion entries for a 20-qubit system. Storing this matrix demands tens of terabytes of memory, which is only available as distributed memory in sizable clusters. Analogously, a quantum system with more

[1] A mixed state is the most general way to express a quantum state.

than 30 qubits would demand $100\times$ more memory than the one present in the world's fastest supercomputer.

To alleviate these challenges, we study the following distributed optimization problem to be jointly solved over a set of M machines:

$$\min_{X \in \mathbb{R}^{d \times d}} \left\{ f(X) = \frac{1}{M} \sum_{i=1}^{M} f_i(X) \right\}$$
$$\text{subject to } X \succeq 0, \ \operatorname{rank}(X) \leq r. \tag{2}$$

In (2), $f_i(X) := \mathbb{E}_{j \sim \mathcal{D}_i}[f_i^j(X)]$, and $f_i^j(X)$ is the loss function evaluated at the j-th observation of the locally stored dataset of machine i, which follows the distribution $\mathcal{D}_i$.[2] The function $f_i : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a (restricted) strongly convex/smooth differentiable function, and $X \succeq 0$ is the set of positive semi-definite matrices with $\operatorname{rank}(X) \leq r$.[3]

To solve (2), we introduce the *Local Stochastic Factored Gradient Descent* (Local SFGD) algorithm, and prove its convergence. To the best of our knowledge, this is the first work that studies Local SGD in the non-convex factorized objective, and provides convergence in terms of the distance to the optimal model parameter. Our contributions can be summarized as:

- We introduce a distributed problem setup for QST as an instance of (2).
- We propose Local SFGD, a distributed algorithm that uses matrix factorization and utilizes local stochastic gradient steps for the minimization of a non-convex function.
- We provide local convergence guarantees for Local SFGD for restricted strongly convex/smooth losses, which is of independent interest, and subsumes the QST problem as a special case.
- We corroborate our theoretical findings with numerical simulations of QST for the Greenberger-Horne-Zeilinger (GHZ) state.

This letter is organized as follows. Section II reviews the QST protocol, and sets up the non-convex distributed objective. Section III introduces the Local SFGD algorithm, followed by Section IV where we provide the main theoretical results along with the proofs. Lastly, in Section V, we use Local SFGD for the reconstruction of the GHZ state.

II. PRELIMINARIES

Classically, the sample complexity m for reconstructing $\rho^\star \in \mathbb{C}^{d \times d}$ is $O(d^2)$, where d itself grows exponentially with n. To address such large sample complexity requirements, we use low-rankness as prior, as many lab-constructed density matrices have low-rank structure, including the maximally-entangled Greenberger-Horne-Zeilinger (GHZ) state [6]. While the low-rank constraint is non-convex, it provides a significant reduction in the sample complexity. Under appropriate assumptions, a rank-r density matrix can be reconstructed with $m = O(r \cdot d \cdot \operatorname{poly} \log(d))$ measurements, instead of $m = O(d^2)$ [7].

---

[2]We assume the homogeneous data case where $\mathcal{D}_i = \mathcal{D}$ for all i.

[3]We provide theory for the real case; extensions to complex domains can be obtained with complex conversions and Wirtinger derivatives [5].

We propose to solve a factorized version of (1) to efficiently handle its low-rank constraint, following [8], [9]:

$$\min_{U \in \mathbb{C}^{d \times r}} G(U) := F(UU^\dagger) = \frac{1}{2m} \|\mathcal{A}(UU^\dagger) - y\|_2^2. \tag{3}$$

In (3), we parametrize the low-rank density matrix ρ by its factor $U \in \mathbb{C}^{d \times r}$. By rewriting $\rho = UU^\dagger$, both the PSD and the low-rank constraints are automatically satisfied, leading to the unconstrained non-convex formulation in (3). Moreover, working in the factored space improves time and space complexities [8]–[10]. However, even with the reduced sample complexity $m = O(r \cdot d \cdot \operatorname{poly} \log(d))$, linear dependency on $d = 2^n$ makes computation infeasible, e.g., for $n = 20$ and rank $r = 100$, the reduced sample complexity still reaches 2.02×10^{10}.

To handle this explosion of data, we consider the setting where the measurements $y \in \mathbb{R}^m$ and the sensing matrices $\mathcal{A} : \mathbb{C}^{d \times d} \to \mathbb{R}^m$ from a central quantum computer *are locally stored across M different classical machines*. These classical machines perform some local operations based on their local data, and communicate back and forth with the central quantum server to reconstruct a density matrix.

The *distributed* QST problem can be written as:

$$\min_{U \in \mathbb{C}^{d \times r}} \left\{ g(U) = \frac{1}{M} \sum_{i=1}^{M} g_i(U) \right\},$$
$$\text{where } g_i(U) := \mathbb{E}_{j \sim \mathcal{D}_i} \|\mathcal{A}_i^j(UU^\dagger) - y_i^j\|_2^2, \tag{4}$$

with j being a random variable that follows a distribution $\mathcal{D}_i$ for machine i. In the next section, we introduce our approach to solve (4), which can be more generally applied to (2).

III. ALGORITHMS

We now introduce the *Local Stochastic Factored Gradient Descent* (Local SFGD) algorithm. We review the *Factored Gradient Descent* (FGD) algorithm [8]–[10] and its stochastic variant [11], on which the Local SFGD is based.

⋄ **Factored Gradient Descent (FGD)**. A common approach to solve the factorized non-convex objective in (3) in *centralized* settings is to use gradient descent on the factor U:

$$U_{t+1} = U_t - \eta_t \nabla G(U_t) = U_t - \eta_t \nabla F(U_t U_t^\dagger) \cdot U_t$$
$$= U_t - \frac{\eta_t}{m} \left(\sum_{k=1}^{m} \left\{ \operatorname{Tr}(A_k U_t U_t^\dagger) - y_k \right\} A_k \right) \cdot U_t, \tag{5}$$

where $\eta_t > 0$ is the step size. From (5), we can see that a pass over full data is required to compute the gradient on every iteration. This can be computationally challenging or even infeasible when m is large, which is almost always the case for QST, even for moderate number of qubits n.

⋄ **Stochastic Factored Gradient Descent (SFGD)**. A simple and effective way to mitigate this burden is to use the *Stochastic Factored Gradient Descent* (SFGD), which replaces the true gradient ∇G with an unbiased estimator H. For instance, one can use the following SFGD update:

$$U_{t+1} = U_t - \eta_t \cdot H(U_t)$$
$$= U_t - \frac{\eta_t}{b} \left(\sum_{k=1}^{b} \left\{ \operatorname{Tr}(A_k U_t U_t^\dagger) - y_k \right\} A_k \right) \cdot U_t, \tag{6}$$

which simply uses b measurements instead of $m \gg b$ to approximate ∇G, where the hyperparameter b is the batch size. In [11], the convergence of SFGD was shown for (restricted) strongly convex/smooth functions.

From (6), one can see that SFGD is amenable to parallelization, simply by replacing $H(U_t)$ with an average of the stochastic gradients that are computed independently from the local machines. This scheme is often the state-of-the-art in distributed learning problems [12], [13]. However, it exhibits a major drawback: on every iteration, each machine has to send the local (stochastic) gradient to the server, and receive back the aggregated model parameter. Such communication is much more expensive than—typically about 3 orders of magnitude—the local computations that each machine has to perform [14].

◇ **Our approach: Local SFGD**. There are two main approaches to resolve the aforementioned communication overhead. One is to reduce the number of transmitted bits via gradient compression schemes, such as quantization [15] or sparsification [16]. The other is to increase the amount of local iterations performed on each machine, in order to reduce the total communication rounds. The latter approach is called Local SGD, and was shown to outperform (parallel) SGD in some settings [17]–[20].

In this letter, we introduce the Local SFGD to estimate the low-rank factor of a density matrix over a set of local machines. Although our main application is to solve the distributed QST objective in (4), Local SFGD is more generally applicable to the distributed objective in (2); see Section IV for details. Local SFGD is summarized in Algorithm 1. While there are non-convex results on Local SGD [21], [22], they consider a different problem setting, and only provide convergence in terms of the norm of the gradient. To the best of our knowledge, this is the first work that studies Local SGD in the non-convex factorized objective, and provides convergence in terms of the distance to the optimal model parameters.

Local SFGD produces M sequences in parallel, where M is the number of machines. If a synchronization step happens at time t, i.e., $t = t_p$ for some $p \in \mathbb{N}$, then the local parameters at each machine U_t^i are sent to the central server, and their average is computed (line 6). Otherwise, each machine performs (possibly many iterations of) SFGD without communicating with the central server (line 8). An important metric to consider for Local SFGD is the maximum time interval between two synchronization time steps: $\max_p |t_p - t_{p+1}|$, which we assume is bounded by $h \geq 1$; see also Theorems 1 and 2. If communication happens on every iteration, i.e., $h = 1$, then Algorithm 1 reduces to the (parallel) SFGD in (6).

IV. THE CONVERGENCE OF LOCAL SFGD

We now provide the local convergence guarantees of Local SFGD in Algorithm 1 for *restricted μ-strongly convex/L-smooth objectives*. Similarly to (3), as we parametrize $X = UU^\top$, Problem (2) becomes non-convex:

$$\min_{U \in \mathbb{R}^{d \times r}} \left\{ g(U) = \frac{1}{M} \sum_{i=1}^{M} g_i(U) \right\}, \qquad (8)$$

which now is unconstrained, as both the PSD and the low-rank constraints are automatically satisfied.

Algorithm 1 Local SFGD

1: Set number of iterations $T > 0$, synchronization time steps $t_1, t_2, \ldots$, and initialize $U_0^i = U_0$ as below:

$$U_0^i = \text{SVD}\left(- \sum_{i=1}^{M} \frac{m_i}{m} \nabla f_i(0) \right) \quad \forall i \in [M], \qquad (7)$$

where SVD denotes the singular value decomposition.
2: **for** each round $t = 0, \ldots T$ **do**
3:     **for** in parallel for $i \in [M]$ **do**
4:         Sample j_t uniformly at random from $[m_i]$.
5:         **if** $t = t_p$ for some $p \in \mathbb{N}$ **then**
6:             $U_{t+1}^i = \frac{1}{M} \sum_{i=1}^{M} \left(U_t^i - \eta_t \nabla g_i^{j_t}(U_t^i) \right)$
7:         **else**
8:             $U_{t+1}^i = U_t^i - \eta_t \nabla g_i^{j_t}(U_t^i)$
9:         **end if**
10:     **end for**
11: **end for**
12: **return** $\hat{U}_{T+1} := \frac{1}{M} \sum_{i=1}^{M} U_{T+1}^i$.

We assume f_i is a symmetric function: $f_i(X) = f_i(X^\top)$. Then, the gradient of $g_i(U) = f_i(UU^\top)$ simplifies to:[4]

$$\nabla g_i(U) = \left(\nabla f_i(UU^\top) + \nabla f_i(UU^\top)^\top \right) U = 2 \nabla f_i(UU^\top) U.$$

We now state the key assumptions used in our main results.

Assumption 1: The function f_i is μ-restricted strongly convex and L-restricted smooth. That is, $\forall X, Y \succeq 0$ and $\forall i \in [M]$, it holds that

$$f_i(Y) \geq f_i(X) + \langle \nabla f_i(X), Y - X \rangle + \frac{\mu}{2} \|X - Y\|_F^2, \quad \text{(I-a)}$$

$$\text{and} \quad \|\nabla f_i(X) - \nabla f_i(Y)\|_F \leq L \|X - Y\|_F. \quad \text{(I-b)}$$

Assumption 2: The stochastic gradient ∇g_i^j is unbiased, has a bounded variance, and is bounded in expectation, $\forall i \in [M]$. That is,

$$\mathbb{E}_j \left[\nabla g_i^j(U) \right] = \nabla g_i(U), \qquad \text{(II-a)}$$

$$\mathbb{E}_j \left[\|\nabla g_i^j(U) - \nabla g_i(U)\|_F^2 \right] \leq \sigma^2, \quad \text{and} \qquad \text{(II-b)}$$

$$\mathbb{E}_j \left[\|\nabla g_i^j(U)\|_F^2 \right] \leq G^2, \qquad \text{(II-c)}$$

where j follows a uniform distribution.

Assumptions (I-a) and (I-b) respectively state that μ-strong convexity and L-smoothness hold when we restrict the space of $d \times d$ matrices to the set of PSD matrices. Such assumptions have become standard in optimization analysis, and are significantly weaker than assuming global strong convexity. Importantly, note that we only assume $f_i(X)$ to have such structures—the transformed function $g_i(U)$ in (8) typically does not satisfy restricted strong convexity/smoothness [23].

Assumptions (II-a) and (II-b) respectively imply that the stochastic gradient is unbiased and has a bounded variance, and both are standard assumptions in stochastic optimization [24]. Assumption (II-c) states that the stochastic gradient has a bounded norm, in expectation. This assumption may seem strong when the objective is (unconstrained) strongly convex [25]; however, note that Assumption (I-a) is restricted to PSD matrices, and the original Problem (2) is constrained.

Apart from (8) being non-convex, another difficulty that arises by the parametrization $X = UU^\top$ is that the solution

---

[4]Without loss of generality, we absorb 2 into η_t to use $\nabla g_i(U)$ and $\nabla f_i(UU^\top)U$ interchangeably.

can become non-unique.[5] We can remove this ambiguity by defining the following rotation invariant distance metric.

Definition 1 [10, eq. 3.1]: For any $U, V \in \mathbb{R}^{d \times r}$, let $D(U, V) := \min_{R \in \mathcal{O}} \|U - VR\|_F$, where $\mathcal{O} \subseteq \mathbb{R}^{r \times r}$ is the set of orthonormal matrices such that $R^\top R = \mathbb{I}_{r \times r}$.

Remark 1: Definition 1 regards all $U, V \in \mathbb{R}^{d \times r}$ to be in the same distance such that $D(U, VR) = D(UR, V) = D(U, V)$. Hence, it defines the equivalence classes $\{UR : R^\top R = \mathbb{I}_{r \times r}\}$ and $\{VR : R^\top R = \mathbb{I}_{r \times r}\}$ [23].

A crucial component for our convergence analysis is the following lemma, which replaces the role of (strong) convexity in classical convergence analysis of gradient descent:

Lemma 1 [26, Lemma 14]: Let Assumption 1 hold. Assume that $D^2(U_0^i, U^\star) \leq \frac{\sigma_r(X^\star)}{100 \cdot \kappa \cdot \sigma_1(X^\star)}$, where $\sigma_k(X^\star)$ is the k-th singular value of $X^\star$,[6] and $\kappa := \frac{L}{\mu}$. Then, the following inequality holds:

$$\langle U_t^i - U^\star R^\star, \nabla g_i(U_t^i)\rangle$$
$$\geq \frac{2\eta_t}{3}\|\nabla g_i(U_t^i)\|_F^2 + \frac{3\mu}{20}\sigma_r(X^\star) \cdot D^2(U_t^i, U^\star). \quad (9)$$

Remark 2: The initialization scheme (7) in Algorithm 1 is modified from [26, Th. 11] to distributed version, and satisfies the initialization condition of Lemma 1 for small enough κ; for the QST problem, the Pauli sensing matrices A_k satisfy the Restricted Isometry Property (RIP) [4], [27], implying $\kappa \approx \frac{1+\delta}{1-\delta}$, where $\delta \in (0, 1)$ is the RIP constant. Hence, by using the right prior information (e.g., low-rankness), we can apply compressed sensing results, implying that $X^\star$ is unique and can be recovered exactly [28].

We are now ready to present the main theoretical results. We first show in Theorem 1 that the Local SFGD converges locally at a linear rate to a small neighborhood of the global optimum with a constant step size. Then, in Theorem 2, we show the exact local convergence by using appropriately diminishing step sizes, at the expense of reducing the convergence rate to a sub-linear rate. We first state the following auxiliary lemma that we use in the proof of Theorems 1 and 2.

Lemma 2: Let Assumptions 1 and (II-c) hold. Then, the output of Algorithm 1 with $\max_p |t_{p+1} - t_p| \leq h$ satisfies:

$$\frac{1}{M}\sum_{i=1}^M \mathbb{E}\Big[\|\hat{U}_t - U_t^i\|_F^2\Big] \leq \eta_{t_q}^2 (h-1)^2 G^2, \quad (10)$$

where t_q is the synchronization step immediately before t. The proof follows the same arguments as [17, Lemma 3.3], and hence is omitted.

Theorem 1 (Local Linear Convergence With Constant Step Size): Let Assumptions 1, 2, and the initialization condition of Lemma 1 hold. Moreover, let $\eta_t = \eta < \frac{1}{\alpha}$ for $t \in [0 : T]$ and $\max_p |t_p - t_{p+1}| \leq h$. Then, the output of Algorithm 1 has the following property:

$$\mathbb{E}\Big[D^2(\hat{U}_{T+1}, U^\star)\Big] \leq (1 - \eta\alpha)^{T+1} D^2(\hat{U}_0, U^\star)$$
$$+ \eta\Big(\frac{4(h-1)^2 G^2}{\alpha} + \frac{\sigma^2}{M\alpha}\Big), \quad (11)$$

where $X^\star$ is the optimum of f over the set of PSD matrices such that $\text{rank}(X^\star) = r$, $U^\star$ is such that $X^\star = U^\star U^{\star\top}$, and $\alpha = \frac{3\mu}{10}\sigma_r(X^\star)$ is a global constant.

---

[5]Consider reconstructing $X^\star = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. It can be seen that both $U^\star = [1\ 1]^\top$ and $\tilde{U}^\star = -[1\ 1]^\top$ satisfy $U^\star U^{\star\top} = \tilde{U}^\star \tilde{U}^{\star\top} = X^\star$.

[6]Without loss of generality, singular values are sorted in descending order.

Remark 3: In (11), the expectation is with respect to the previous iterates, $\{\hat{U}_t\}_{t=0}^T$. We make a few remarks about Theorem 1. First, notice the last variance term $\frac{\sigma^2}{M\alpha}$, which disappears in the noiseless case, is reduced by the number of machines M. Second, we assume single-batch is used in the proof; by using batch size $b > 1$, this term can be further divided by b. Last, by plugging in $h = 1$ (i.e., synchronization happens on every iteration), the first variance term disappears, exhibiting similar local linear convergence to SFGD [11].

Proof: Throughout the proof, we use the notations:

$$U_{t+1}^i = U_t^i - \eta_t g_t^i \quad \text{and} \quad \hat{U}_{t+1} = \hat{U}_t - \eta_t g_t,$$

where $\hat{U}_t = \frac{1}{M}\sum_{i=1}^M U_t^i$, i.e., the average across different machines at time t. We denote the stochastic gradient of machine i at time t with $g_t^i := \nabla f_i^{j_t}(U_t^i U_t^{i\top})U_t^i = \nabla g_i^{j_t}(U_t^i)$, and the average of stochastic gradients across machines with $g_t = \frac{1}{M}\sum_{i=1}^M g_t^i$. Finally, we denote $\mathbb{E}[g_t] = \bar{g}_t$.

We first decompose the distance of $D^2(\hat{U}_{t+1}, U^\star)$:

$$D^2(\hat{U}_{t+1}, U^\star) = \min_{R \in \mathcal{O}}\|\hat{U}_{t+1} - U^\star R\|_F^2 \leq \|\hat{U}_{t+1} - U^\star R^\star\|_F^2$$
$$= \|\hat{U}_t - U^\star R^\star - \eta_t \bar{g}_t\|_F^2 + \eta_t^2 \|\bar{g}_t - g_t\|_F^2$$
$$+ 2\eta_t\langle \hat{U}_t - U^\star R^\star - \eta_t \bar{g}_t, \bar{g}_t - g_t\rangle. \quad (12)$$

The first term in (12) can be further decomposed to:

$$\|\hat{U}_t - U^\star R^\star\|_F^2 + \eta_t^2 \|\bar{g}_t\|_F^2 - 2\eta_t\langle \hat{U}_t - U^\star R^\star, \bar{g}_t\rangle. \quad (13)$$

We bound the second and the third terms in (13) separately. For the second term, by Jensen's inequality, we have:

$$\|\bar{g}_t\|_F^2 = \Big\|\frac{1}{M}\sum_{i=1}^M \nabla g_i(U_t^i)\Big\|_F^2 \leq \frac{1}{M}\sum_{i=1}^M \|\nabla g_i(U_t^i)\|_F^2. \quad (14)$$

For the third term, we decompose further to have:

$$\Big\langle \hat{U}_t - U^\star R^\star, \frac{1}{M}\sum_{i=1}^M \nabla g_i(U_t^i)\Big\rangle = \frac{1}{M}\sum_{i=1}^M \langle \hat{U}_t - U_t^i, \nabla g_i(U_t^i)\rangle$$
$$+ \frac{1}{M}\sum_{i=1}^M \langle U_t^i - U^\star R^\star, \nabla g_i(U_t^i)\rangle. \quad (15)$$

We again bound the two terms in (15) separately. Using $\langle A, B\rangle \geq -\frac{\delta}{2}\|A\|_F^2 - \frac{1}{2\delta}\|B\|_F^2$, the first term admits:

$$\frac{1}{M}\sum_{i=1}^M \langle \hat{U}_t - U_t^i, \nabla g_i(U_t^i)\rangle$$
$$\geq \frac{1}{M}\sum_{i=1}^M \Big(-\frac{\delta}{2}\|\hat{U}_t - U_t^i\|_F^2 - \frac{1}{2\delta}\|\nabla g_i(U_t^i)\|_F^2\Big)$$

By Lemma 1, the second term in (15) admits:

$$\frac{1}{M}\sum_{i=1}^M \langle U_t^i - U^\star R^\star, \nabla g_i(U_t^i)\rangle$$
$$\geq \frac{1}{M}\sum_{i=1}^M \frac{2\eta_t}{3}\|\nabla g_i(U_t^i)\|_F^2 + \frac{1}{M}\sum_{i=1}^M \frac{3\mu \cdot \sigma_r(X^\star)}{20}D^2(U_t^i, U^\star)$$
$$\geq \frac{1}{M}\sum_{i=1}^M \frac{2\eta_t}{3}\|\nabla g_i(U_t^i)\|_F^2 + \frac{3\mu \cdot \sigma_r(X^\star)}{20}D^2(\hat{U}_t, U^\star),$$

where we used the convexity of $D^2(\cdot, \cdot)$ in the last inequality.

Combining above two bounds into (15), we have

$$\langle \hat{U}_t - U^\star R^\star, \bar{g}_t \rangle$$
$$\geq \frac{1}{M} \sum_{i=1}^{M} \left(-\frac{\delta}{2} \|\hat{U}_t - U_t^i\|_F^2 - \frac{1}{2\delta} \|\nabla g_i(U_t^i)\|_F^2 \right)$$
$$+ \frac{1}{M} \sum_{i=1}^{M} \frac{2\eta_t}{3} \|\nabla g_i(U_t^i)\|_F^2 + \frac{3\mu \cdot \sigma_r(X^\star)}{20} D^2(\hat{U}_t, U^\star). \quad (16)$$

Substituting (14) and (16) into (13), we have

$$\|\hat{U}_t - U^\star R^\star - \eta_t \bar{g}_t\|_F^2$$
$$\leq \left(1 - \eta_t \cdot \frac{3\mu}{10} \sigma_r(X^\star) \right) \|\hat{U}_t - U^\star R^\star\|_F^2$$
$$+ \frac{1}{M} \sum_{i=1}^{M} \left[\left(\frac{\eta_t}{\delta} - \frac{\eta_t^2}{3} \right) \|\nabla g_i(U_t^i)\|_F^2 + \eta_t \delta \|\hat{U}_t - U_t^i\|_F^2 \right]$$
$$\overset{\delta = 4/\eta_t}{=} (1 - \eta_t \alpha) \|\hat{U}_t - U^\star R^\star\|_F^2$$
$$+ \frac{1}{M} \sum_{i=1}^{M} \left[\eta_t \left(\frac{\eta_t}{4} - \frac{\eta_t}{3} \right) \|\nabla g_i(U_t^i)\|_F^2 + 4 \|\hat{U}_t - U_t^i\|_F^2 \right]$$
$$\leq (1 - \eta_t \alpha) \|\hat{U}_t - U^\star R^\star\|_F^2 + \frac{4}{M} \sum_{i=1}^{M} \|\hat{U}_t - U_t^i\|_F^2, \quad (17)$$

where in the equality we defined $\alpha := \frac{3\mu}{10} \sigma_r(X^\star)$, and in the last inequality we used that $\frac{\eta_t}{4} - \frac{\eta_t}{3} < 0$.

Substituting (17) into (12) and taking expectations conditional on the previous iterates, and using $\mathbb{E}[g_t] = \bar{g}_t$, we get

$$\mathbb{E}[D^2(\hat{U}_{t+1}, U^\star)] = \|\hat{U}_t - U^\star R^\star - \eta_t \bar{g}_t\|_F^2 + \eta_t^2 \mathbb{E}[\|\bar{g}_t - g_t\|_F^2]$$
$$\overset{(17)}{\leq} (1 - \eta_t \alpha) \|\hat{U}_t - U^\star R^\star\|_F^2$$
$$+ \frac{4}{M} \sum_{i=1}^{M} \mathbb{E}\left[\|\hat{U}_t - U_t^i\|_F^2 \right] + \eta_t^2 \mathbb{E}\left[\|\bar{g}_t - g_t\|_F^2 \right]$$
$$\overset{(10)}{\leq} (1 - \eta_t \alpha) \|\hat{U}_t - U^\star R^\star\|_F^2$$
$$+ 4\gamma^2 \eta_t^2 (h-1)^2 G^2 + \eta_t^2 \mathbb{E}[\|\bar{g}_t - g_t\|_F^2],$$

where the last inequality is by Lemma 2 with the fact that $\eta_{t_q}^2 \leq \gamma^2 \eta_t^2$ for some constant γ, which holds for proper choice of $\{\eta_t\}_{t=0}^{T}$. We further have:

$$\mathbb{E}[\|\bar{g}_t - g_t\|_F^2] \leq \frac{1}{M^2} \sum_{i=1}^{M} \mathbb{E}[\|\nabla g_i^{j_t}(U_t^i) - \nabla g_i(U_t^i)\|_F^2] \overset{(\text{II-b})}{\leq} \frac{\sigma^2}{M},$$

where we used $\text{Var}(\sum_{m=1}^{M} X_m) = \sum_{m=1}^{M} \text{Var}(X_m)$ for independent random variables.

We now arrive at the iteration invariant bound:

$$\mathbb{E}[D^2(\hat{U}_{t+1}, U^\star)]$$
$$\leq (1 - \eta_t \alpha) D^2(\hat{U}_t, U^\star) + \eta_t^2 \left(4\gamma^2 (h-1)^2 G^2 + \frac{\sigma^2}{M} \right). \quad (18)$$

Lastly, unfolding (18) for T iterations, and using $\gamma = 1$ for the constant step size[7] $\eta_t = \eta$ for $t \in [0 : T]$ as well as

---

[7]For the decreasing step sizes in Theorem 2, it holds that $\gamma \leq \frac{(h+2)}{2}$.

---

the fact that $\sum_{t=0}^{T} (1 - \eta\alpha)^t \leq \sum_{t=0}^{\infty} (1 - \eta\alpha)^t = \frac{1}{\eta\alpha}$, we obtain:

$$\mathbb{E}[D^2(\hat{U}_{T+1}, U^\star)] \leq (1 - \eta\alpha)^{T+1} D^2(\hat{U}_0, U^\star)$$
$$+ \eta \left(\frac{4(h-1)^2 G^2}{\alpha} + \frac{\sigma^2}{M\alpha} \right),$$

which completes the proof. ∎

Theorem 2 (Local Sub-Linear Convergence With Diminishing Step Sizes): Let Assumptions 1, 2, and the initialization condition of Lemma 1 hold. Moreover, let $\eta_t = \frac{2}{\alpha(t+2)}$ for $t \in [0 : T]$ and $\max_p |t_p - t_{p+1}| \leq h$. Then, the output of Algorithm 1 has the following property:

$$\mathbb{E}\left[D^2(\hat{U}_{T+1}, U^\star) \right] \leq \frac{4C}{\alpha(T+3)}, \quad (19)$$

where $X^\star$ is the optimum such that $\text{rank}(X^\star) = r$, $U^\star$ is such that $X^\star = U^\star U^{\star\top}$, and $\alpha = \frac{3\mu}{10} \sigma_r(X^\star)$ and $C = (h-1)^2(h+2)^2 G^2 + \frac{\sigma^2}{M}$ are global constants.

Proof: We claim the following, and prove by induction:

$$D^2(\hat{U}_t, U^\star) \leq \frac{4C}{\alpha^2(t+2)}, \quad \text{with} \quad \eta_t = \frac{2}{\alpha(t+2)}. \quad (20)$$

We start from the iteration invariant bound in (18):

$$\mathbb{E}\left[D^2(\hat{U}_{t+1}, U^\star) \right] \leq (1 - \eta_t \alpha) D^2(\hat{U}_t, U^\star) + \eta_t^2 \cdot C.$$

For the base case $t = 0$, we have

$$\mathbb{E}\left[D^2(\hat{U}_1, U^\star) \right] \leq (1 - \eta_0 \alpha) D^2(\hat{U}_0, U^\star) + \eta_0^2 \cdot C$$
$$= \left(1 - \frac{1}{\alpha} \cdot \alpha \right) D^2(\hat{U}_0, U^\star) + \frac{C}{\alpha^2} = \frac{C}{\alpha^2} \leq \frac{4C}{3\alpha^2}.$$

Now, we proceed to the inductive step. Assuming (20) holds for the time step t, we want to prove the same holds for the time step $t + 1$. Starting from (18) again, we have:

$$\mathbb{E}\left[D^2(\hat{U}_{t+1}, U^\star) \right] \leq (1 - \eta_t \alpha) D^2(\hat{U}_t, U^\star) + \eta_t^2 \cdot C$$
$$\overset{(20)}{\leq} \left(1 - \frac{2}{t+2} \right) \cdot \frac{4C}{\alpha^2(t+2)} + \frac{4C}{\alpha^2(t+2)^2} = 4C \cdot \frac{t+1}{\alpha^2(t+2)} \cdot \frac{1}{t+2}$$
$$\leq 4C \cdot \frac{t+2}{\alpha^2(t+3)} \cdot \frac{1}{t+2} = \frac{4C}{\alpha^2(t+3)},$$

where in the last inequality we used the fact that $\frac{t+1}{t+2} \leq \frac{t+2}{t+3}$. This completes the proof. ∎

V. NUMERICAL RESULTS

We use the Local SFGD to reconstruct the Greenberger-Horne-Zeilinger (GHZ) state, using simulated measurement data from Qiskit. GHZ state is known as *maximally entangled* quantum state [6], meaning that it exhibits the maximal inter-particle correlation, which does not exist in the classical mechanics. We are interested in: (*i*) how the number of local steps affect the accuracy defined as $\varepsilon = \|\hat{U}_t \hat{U}_t^\top - \rho_{\text{ghz}}^\star\|_F^2$, where $\rho_{\text{ghz}}^\star = U^\star U^{\star\dagger}$ is the true density matrix for the GHZ state; and (*ii*) the scalability of the distributed setup for various number of classical machines M.

In Fig. 1 (Top), we first fix the number of machines $M = 10$ and the number of total synchronization steps to be 100, and vary the number of local iterations between two synchronization steps, i.e., $h \in \{1, 10, 25, 50, 100, 200\}$. We use constant step size $\eta = 1$ for all h. Increasing h, i.e., each distributed
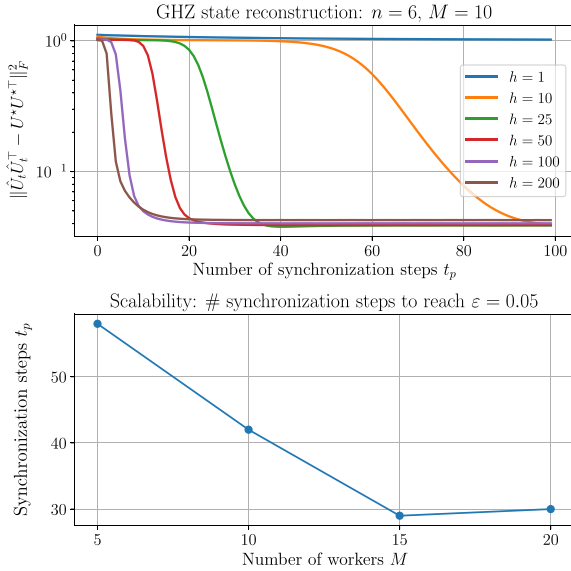
Fig. 1. Top: Convergence speed as a function of number of synchronization steps t_p for various number of local iterations. Bottom: number of synchronization steps to reach $\varepsilon \leq 0.05$ as a function of number of workers M. The batch size $b = 50$ is used for all cases.

machine performing more local iterations, leads to faster convergence in terms of the synchronization steps. Notably, the speed up gets marginal: e.g., there is not much difference between $h = 100$ and $h = 200$, indicating there is an "optimal" h that leads to the biggest reduction in the number of synchronization steps. Further, one can notice that higher h leads to slightly worse final accuracy—this is consistent with (11) in Theorem 1, where the first variance term that depends on G^2 disappears with $h = 1$. Finally, note that ε does not decrease below certain level due to the inherent finite sampling error of quantum measurements [29].

In Fig. 1 (Bottom), we plot the number of synchronization steps to reach $\varepsilon \leq 0.05$, while fixing $h = 20$. We vary the number of workers $M \in \{5, 10, 15, 20\}$, where each machine gets 200 measurements. There is a significant speed up from $M = 5$ to $M = 15$, while for $M = 20$, it took one more synchronization step compared to $M = 15$, which is likely due to the stochasticity of SFGD within each machine.

VI. CONCLUSION AND FUTURE WORK

In this letter, we introduced a distributed problem set up for QST as an instance of a general distributed optimization problem with PSD/low-rank constraints. We proposed the Local SFGD, a distributed non-convex algorithm that utilizes local steps at each distributed worker to estimate the low-rank factor of a density matrix. We proved the local convergence of Local SFGD for restricted strongly convex/smooth objectives, which can be of independent interest. For future work, extension to the heterogeneous data case as well as the decentralized case with various topologies can be investigated.

REFERENCES

[1] J. W. Park, Z. Z. Yan, H. Loh, S. A. Will, and M. W. Zwierlein, "Second-scale nuclear spin coherence time of ultracold ^{23}Na^{40}K molecules," *Science*, vol. 357, pp. 372–375, 2017.

[2] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, Aug. 2018.

[3] J. B. Altepeter, E. R. Jeffrey, and P. G. Kwiat, "Photonic state tomography," *Adv. Atomic Mol. Opt. Phys.*, vol. 52, pp. 105–159, Jan. 2005.

[4] Y.-K. Liu, "Universal low-rank matrix recovery from pauli measurements," in *Advances in NeurIPS*, vol. 24. Red Hook, NY, USA: Curran, 2011.

[5] R. C. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, vol. 368. Providence, RI, USA: Amer. Math. Soc., 2009.

[6] Y. Zhao, R. Zhang, W. Chen, X.-B. Wang, and J. Hu, "Creation of Greenberger-Horne-Zeilinger states with thousands of atoms by entanglement amplification," *NPJ Quantum Inf.*, vol. 7, no. 1, pp. 1–6, 2021.

[7] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, no. 15, 2010, Art. no. 150401.

[8] A. Kyrillidis, A. Kalev, D. Park, S. Bhojanapalli, C. Caramanis, and S. Sanghavi, "Provable compressed sensing quantum state tomography via non-convex methods," *NPJ Quantum Inf.*, vol. 4, no. 1, pp. 1–7, 2018.

[9] J. L. Kim, G. Kollias, A. Kalev, K. X. Wei, and A. Kyrillidis, "Fast quantum state reconstruction via accelerated non-convex programming," 2021, *arXiv:2104.07006*.

[10] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. ICML*, 2016, pp. 964–973.

[11] J. Zeng, K. Ma, and Y. Yao, "On global linear convergence in stochastic nonconvex optimization for semidefinite programming," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4261–4275, Aug. 2019.

[12] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[13] P. Goyal *et al.*, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," 2017, *arXiv:1706.02677*.

[14] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Math. Program.*, vol. 180, no. 1, pp. 237–284, 2020.

[15] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in NeurIPS*, vol. 30. Red Hook, NY, USA: Curran, 2017.

[16] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks," *J. Mach. Learn. Res.*, vol. 22, no. 241, pp. 1–124, 2021.

[17] S. U. Stich, "Local SGD converges fast and communicates little," 2018, *arXiv:1805.09767*.

[18] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. R. Cadambe, "Local SGD with periodic averaging: Tighter analysis and adaptive synchronization," in *NeurIPS*, vol. 32. Red Hook, NY, USA: Curran, 2019.

[19] B. Woodworth *et al.*, "Is local SGD better than minibatch SGD?" in *Proc. ICML*, 2020, pp. 10334–10343.

[20] A. Khaled, K. Mishchenko, and P. Richtarik, "Tighter theory for local SGD on identical and heterogeneous data," in *Proc. AISTATS*, Jun. 2020, pp. 4519–4529.

[21] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," 2018, *arXiv:1808.07576*.

[22] F. Zhou and G. Cong, "On the convergence properties of a K-step averaging stochastic gradient descent algorithm for nonconvex optimization," 2017, *arXiv:1708.01012*.

[23] D. Zhou, Y. Cao, and Q. Gu, "Accelerated factored gradient descent for low-rank matrix factorization," in*Proc. AISTATS*, vol. 108, 2020, p. 10.

[24] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[25] L. M. Nguyen, P. H. Nguyen, M. V. Dijk, P. Richtárik, K. Scheinberg, and M. Takáč, "SGD and Hogwild! convergence without the bounded gradients assumption," in *Proc. ICML*, 2018, pp. 3750–3758.

[26] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Proc. Conf. Learn. Theory*, 2016, pp. 530–582.

[27] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[28] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[29] O. Crawford, B. van Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley, "Efficient quantum measurement of pauli operators in the presence of finite sampling error," *Quantum*, vol. 5, p. 385, Jan. 2021.