Local Stochastic Factored Gradient Descent for Distributed Quantum State Tomography

Junhyung Lyle Kim (Rice CS), Mohammad Taha Toghani (Rice ECE),
CésarA.Uribe (Rice ECE), Anastasios Kyrillidis (Rice CS)

[QST objective]

- A quantum state can be represented by a density matrix ρ which is a complex, positive semi-definite (PSD) matrix with unit trace
- Estimating ρ, given the measurement data, is the goal of QST
- The density matrix of an n-qubit mixed state 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 p_k is the probability of finding ρ in the pure state Ψ_k.

- Given theses 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 $d = 2^n$:

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

- $\mathscr{A} : \mathbb{C}^{2^n\times 2^n} \to \mathbb{R}^m$ is the linear sensing map such that $\mathscr{A}(\rho)_k = \text{Tr}(A_k\rho)$ for $k = 1,\dots,m$ (the Born rule)

[Motivation of low-rank prior]

- Classically (without low-rank prior), the sample complexity m for reconstructing $\rho^\star \in \mathbb{C}^{d\times d}$ is $O(d^2)$
- [Gross et al., 2010] proved that a rank-r density matrix can be reconstructed with $m = O(r \cdot d \cdot \text{poly}\log(d))$ measurements instead
- However, low-rankness is a non-convex constraint, which is tricky to handle

[Modified QST objective]

- By rewriting $\rho = UU^\dagger$, both the PSD and the low-rank constraints are automatically satisfied, leading to the following unconstrained non-convex formulation:

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

- Even with the reduced sample complexity $m = O(r \cdot d \cdot \text{poly}\log(d))$, its linear dependency on $d = 2^n$ is still prohibitively expensive
- E.g., for $n = 20$ and rank $r = 100$, the reduced sample complexity still reaches 2.02×10^{10}

[Distributed QST objective]

- To handle the explosion of data, we consider the setting where the measurements $y \in \mathbb{R}^m$ and the sensing matrices $\mathscr{A} : \mathbb{C}^{d\times d} \to \mathbb{R}^m$ 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.
- The distributed QST problem is:

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

- with j being a random variable that follows a distribution $\mathscr{D}_i$ for machine i.

[Algorithm]

Algorithm 1 Local SFGD

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

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

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

[Assumptions]

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*

$$f_i(Y) \ge f_i(X) + \langle \nabla f_i(X), Y - X\rangle + \frac{\mu}{2}\|X - Y\|_F^2, \quad \text{(I-a)}$$
$$\text{and} \quad \|\nabla f_i(X) - \nabla f_i(Y)\|_F \le L\|X - Y\|_F. \quad \text{(I-b)}$$

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,*

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

where j follows a uniform distribution.

Definition 1 (Eq. (3.1) in [10]). *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 $R^\top R = \mathbb{I}_{r\times r}$.*

Lemma 1 (Lemma 14 in [26]). *Let Assumption 1 hold. Assume that $D^2(U_0^i, U^\star) \le \frac{\sigma_r(X^\star)}{100\cdot\kappa\cdot\sigma_1(X^\star)}$, 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:*

$$\langle U_t^i - U^\star R^\star, \nabla g_i(U_t^i)\rangle \ge \frac{2\eta_t}{3}\|\nabla g_i(U_t^i)\|_F^2 + \frac{3\mu}{20}\sigma_r(X^\star)\cdot D^2(U_t^i, U^\star). \qquad (11)$$

[Constant step size]

Theorem 2 (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 $\max_p |t_p - t_{p+1}| \le h$. Then, the output of Algorithm 1 has the following property:*

$$\mathbb{E}\left[D^2(\hat{U}_{T+1}, U^\star)\right] \le (1-\eta\alpha)^{T+1} D^2(\hat{U}_0, U^\star) + \eta\left(\frac{(h-1)^2G^2}{\alpha} + \frac{\sigma^2}{M\alpha}\right), \qquad (12)$$

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.

- The last variance term $\sigma^2/(M\alpha)$, which disappears in the noiseless case, is reduced by the number of machines M
- Above result assumes a single-batch is used; by using batch size $b > 1$, this term can be further divided by b
- By plugging in $h = 1$ (i.e., synchronization happens on every iteration), the first variance term disappears, exhibiting similar local linear convergence to SFGD.

[Diminishing step sizes]

Theorem 4 (Local sub-linear convergence with diminishing step size). *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 $\max_p |t_p - t_{p+1}| \le h$. Then, the output of Algorithm 1 has the following property:*

$$\mathbb{E}\left[D^2(\hat{U}_{T+1}, U^\star)\right] \le \frac{4C}{\alpha(T+3)}, \qquad (21)$$

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 $C = 4(h-1)^2G^2 + \frac{\sigma^2}{M}$ are global constants.

- We can prove the exact convergence at the cost of slowing down the convergence rate to sub-linear rate by using appropriately diminishing step sizes

[Experiments]

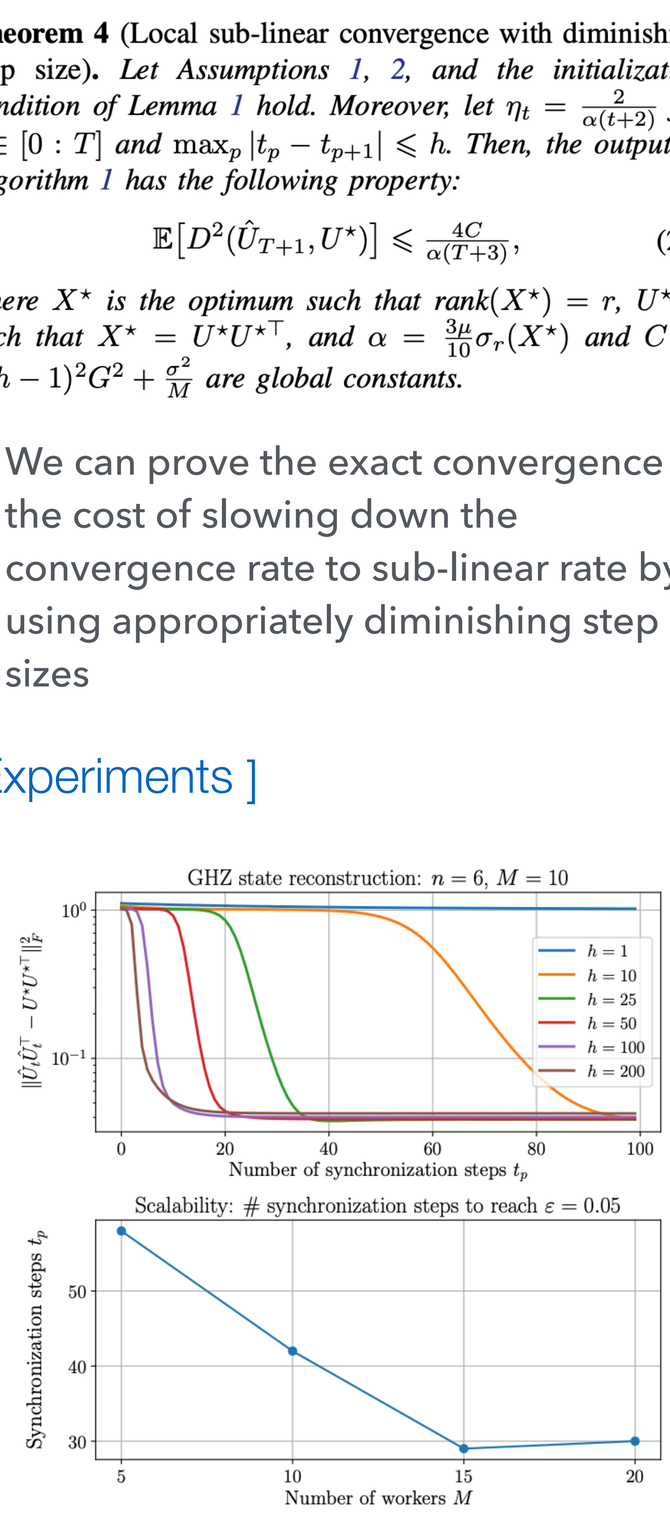

Fig. 1. Top: Convergence speed as a function of number of synchronization steps t_p for various number of local iterations. Bottom: number of synchronization steps to reach $\varepsilon \le 0.05$ as a function of number of workers M. The batch size $b = 50$ is used for all cases.