

Fast quantum state reconstruction via accelerated non-convex programming

Junhyung Lyle Kim¹, George Kollias², Amir Kaley³, Ken X. Wei², Anastasios Kyrillidis^{1*}

¹ Computer Science, Rice University, Houston, TX 77098, USA

² IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

³ Information Sciences Institute, University of Southern California, Arlington, VA 22203, USA

April 15, 2021

Abstract

We propose a new quantum state reconstruction method that combines ideas from compressed sensing, non-convex optimization, and acceleration methods. The algorithm, called Momentum-Inspired Factored Gradient Descent (MiFGD), extends the applicability of quantum tomography for larger systems. Despite being a non-convex method, MiFGD converges *provably* to the true density matrix at a linear rate, in the absence of experimental and statistical noise, and under common assumptions. With this manuscript, we present the method, prove its convergence property and provide Frobenius norm bound guarantees with respect to the true density matrix. From a practical point of view, we benchmark the algorithm performance with respect to other existing methods, in both synthetic and real experiments performed on an IBM's quantum processing unit. We find that the proposed algorithm performs orders of magnitude faster than state of the art approaches, with the same or better accuracy. In both synthetic and real experiments, we observed accurate and robust reconstruction, despite experimental and statistical noise in the tomographic data. Finally, we provide a ready-to-use code for state tomography of multi-qubit systems.

Introduction

Quantum tomography is one of the main procedures to identify the nature of imperfections and deviations in quantum processing unit (QPU) implementation [7, 25]. Generally, quantum tomography is composed of two main parts: *i*) measuring the quantum system, and *ii*) analyzing the measurement data to obtain an estimation of the density matrix (in the case of state tomography [7]), or of the quantum process (in the case of process tomography [63]). In this manuscript, we focus on the case of state tomography.

As the number of free parameters that define quantum states and processes scale exponentially with the number of subsystems, generally quantum tomography is a non-scalable protocol [36]. In particular, quantum state tomography (QST) suffers from two bottlenecks related to its two main parts. The first concerns with the large data one needs to collect to perform tomography; the second concerns with numerically searching in an exponentially large space for a density matrix that is consistent with the data.

There have been various approaches over the years to improve the scalability of QST, as compared to full QST [90, 45, 9]. Focusing on the data collection bottleneck, to reduce the resources required, prior information about the unknown quantum state is often assumed. For example, in compressed sensing QST [36, 46], it is assumed that the density matrix of the system is low-rank. In neural network QST [86, 10, 87], one assumes real and positive wavefunctions, which occupy a restricted place in the landscape of quantum states. Extensions of neural networks to complex wave-functions, or the ability to represent density matrices of mixed states, have been further considered in the literature, after proper reparameterization of the Restricted Boltzmann machines [86]. The prior information considered in these cases is that they are characterized by structured quantum states, which is the reason for the very high performances of neural

*Corresponding author at anastasios@rice.edu

network QST [86].¹ Similarly, in matrix-product-state tomography [21, 55], one assumes that the state-to-be-estimated can be represented with low bond-dimension matrix-product state.

Focusing on the computational bottleneck, several works introduce sophisticated numerical methods to improve the efficiency of QST. Particularly, variations of gradient decent convex solvers—e.g., [33, 14, 79, 41]—are time-efficient in idealized (synthetic) scenarios [41], and only after a proper distributed system design [39]. The problem is that achieving such results seems to require utilizing special-purpose hardware (like graphical-processing units). Thus, going beyond current capabilities requires novel methods that efficiently search in the space of density matrices under more realistic scenarios. Importantly, such numerical methods should come with guarantees on their performance and convergence.

The setup we consider here is that of an n -qubit state, under the prior assumption that the state is close to a pure state, and thus its density matrix is of low-rank. This assumption is justified by state-of-the-art experiments, where our aim is to manipulate the pure states by unitary maps. From a theoretical perspective, the low-rank assumption means that we can use compressed sensing techniques, which allow the recovery of the density matrix from relatively few measurement data [60].

Indeed, by now, compressed sensing QST is widely used for estimating highly-pure quantum states, e.g., [77, 51, 26, 36]. However, compressed sensing QST usually relies on convex optimization for the estimation part [46]; this limits the applicability to relatively small system sizes [36]. On the other hand, non-convex optimization can preform much faster than its convex counterpart [54]. Although non-convex optimization typically lacks convergence guarantees, it was recently shown that one can formulate compressed sensing QST as a non-convex problem and solve it with rigorous convergence guarantees (under certain but generic conditions), allowing state estimation of larger system sizes [54].

Following the non-convex path, we introduce a new algorithm to the toolbox of QST—the Momentum-Inspired Factored Gradient Descent (MiFGD). Our approach combines ideas from compressed sensing, non-convex optimization, and acceleration techniques, to allow pushing QST beyond current capabilities. MiFGD includes acceleration motions per iteration, that non-trivially complicate theoretical justification of its performance in non-convex scenarios, but nevertheless boosts the practical efficiency of the algorithm.

The contributions of the paper are summarized as follows:

- i) We prove that the non-convex MiFGD algorithm has indeed favorable *linear* convergence rate, in terms of iterate distance, in the noiseless measurement data case and under common assumptions.
- ii) We provide QST results based on real data from IBM’s quantum computers up to 8-qubits, contributing to recent efforts on testing QST algorithms in real quantum data [77]. Our synthetic examples scale up to 12-qubits effortlessly, leaving the space for an efficient, hardware-aware implementation open for future work.
- iii) We show in practice that MiFGD allows faster estimation of quantum states compared to state-of-the-art convex and non-convex algorithms, including recent deep learning approaches [86, 10, 87, 29], even in the presence of statistical noise in the measurement data.
- iv) We exploit parallel computations in MiFGD by extending its implementation to enable efficient, parallel execution over shared and distributed memory systems. This way, we experimentally showcase the scalability of this work, which is particularly critical for tackling larger quantum system sizes.
- v) We provide implementation of our approach, compatible to the open-source software Qiskit [75], at <https://github.com/gidiko/MiFGD>.

Results

Setup

We consider the estimation of a low-rank density matrix $\rho^* \in \mathbb{C}^{d \times d}$ on an n -qubit Hilbert space, $d = 2^n$, through the following ℓ_2 -norm reconstruction objective:

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

¹[86] considers also the case of a completely unstructured case and test the limitation of this technique, which does not perform as expected due to lack of structure.

Here, $y \in \mathbb{R}^m$ is the measured data² (observations), and $\mathcal{A}(\cdot) : \mathbb{C}^{d \times d} \rightarrow \mathbb{R}^m$ is the linear sensing map, where $m \ll d^2$. The sensing map relates the density matrix ρ^* to (expected, noiseless) observations through the Born rule, $(\mathcal{A}(\rho))_i = \text{Tr}(A_i \rho)$, where $A_i \in \mathbb{C}^{d \times d}$, $i = 1, \dots, m$, are matrices closely related to the measured observable or the POVM elements of appropriate dimensions. For concreteness, we focus on the least squares objective function.³ The constraint that a density matrix is a non-negative matrix, $\rho \succeq 0$, is a convex constraint. In contrast, the constraint on its rank, $\text{rank}(\rho) \leq r$, is a non-convex constraint that promotes a low-rank solution. Following compressed sensing QST results [46], the constraint $\text{Tr}(\rho) = 1$ (that should be satisfied, by definition, by any density matrix) can be ignored, without affecting the scaling of the precision of the final estimation.

A pivotal assumption is that the linear map \mathcal{A} satisfies the *restricted isometry property*:

Definition 1 (Restricted Isometry Property (RIP) [76]). *A linear operator $\mathcal{A} : \mathbb{C}^{d \times d} \rightarrow \mathbb{R}^m$ satisfies the RIP on rank- r matrices, with parameter $\delta_r \in (0, 1)$, if the following holds for any rank- r matrix $X \in \mathbb{C}^{d \times d}$, with high probability:*

$$(1 - \delta_r) \cdot \|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1 + \delta_r) \cdot \|X\|_F^2. \quad (2)$$

Such maps (almost) preserve the Frobenius norm of low-rank matrices, and, as an extension, of low-rank Hermitian matrices. The intuition behind RIP is that $\mathcal{A}(\cdot)$ behaves as almost a bijection between the subspaces $\mathbb{C}^{d \times d}$ and \mathbb{R}^m , when we focus on low rank matrices.

Following recent works [54], instead of solving Eq. (1), we propose to solve a factorized version of it:

$$\min_{U \in \mathbb{C}^{d \times r}} \frac{1}{2} \|\mathcal{A}(UU^\dagger) - y\|_2^2, \quad (3)$$

where $U^\dagger \in \mathbb{C}^{r \times d}$ denotes the adjoint of U . The motivation for this reformulation is as follows: instead of representing the density matrix ρ as a $d \times d$ Hermitian matrix, and imposing the low-rank constraint as in Eq. (1), we work in a space where low-rank density matrices are represented through factors $U \in \mathbb{C}^{d \times r}$. The low-rankness of ρ is enforced through the factorization of the density matrix into a outer product of such a rectangular matrix representation $U \in \mathbb{C}^{d \times r}$ with its Hermitian conjugate, where $d \gg r$. By rewriting $\rho = UU^\dagger$, for $U \in \mathbb{C}^{d \times r}$, both the PSD constraint ($\rho \succeq 0$) and the low-rankness constraint ($\text{rank}(\rho) \leq r$) are directly satisfied, leading to the non-convex formulation (3). Working in the factored space was shown [54, 73, 72, 88, 95, 96] to improve time and space complexities.

A common approach to solve (3) is to use gradient descent on the parameter U , with iterates generated by the rule:⁴

$$U_{i+1} = U_i - \eta \nabla f(U_i U_i^\dagger) \cdot U_i \quad (4)$$

$$= U_i - \eta \mathcal{A}^\dagger \left(\mathcal{A}(U_i U_i^\dagger) - y \right) \cdot U_i. \quad (5)$$

Here, $U_i \in \mathbb{C}^{d \times r}$, $\forall i$. The operator $\mathcal{A}^\dagger : \mathbb{R}^m \rightarrow \mathbb{C}^{d \times d}$ is the adjoint of \mathcal{A} , defined as $\mathcal{A}^\dagger(x) = \sum_{i=1}^m x_i A_i$, for $x \in \mathbb{R}^m$. The hyperparameter $\eta > 0$ is the step size. This algorithm has been studied in [12, 96, 88, 74, 30, 40]. We will refer to the above iteration as the *Factored Gradient Descent* (FGD) algorithm, as in [73]. In what follows, we will study the *MiFGD* algorithm, a momentum-inspired factored gradient descent.

Momentum-Inspired Factored Gradient Descent and Main Results

The *MiFGD* algorithm is given in the Methods section. It is a two-step variant of FGD:

$$U_{i+1} = Z_i - \eta \mathcal{A}^\dagger \left(\mathcal{A}(Z_i Z_i^\dagger) - y \right) \cdot Z_i, \quad (6)$$

$$Z_{i+1} = U_{i+1} + \mu (U_{i+1} - U_i). \quad (7)$$

²Specific description on how y is generated and what it represents will follow.

³Our results rely on standard optimization assumptions (restricted smoothness and restricted strong convexity assumptions [64]).

⁴We assume cases where $\nabla f(\cdot) = \nabla f(\cdot)^\dagger$. If this does not hold, the theory still holds by carrying around $\nabla f(\cdot) + \nabla f(\cdot)^\dagger$ instead of just $\nabla f(\cdot)$, after proper scaling.

Here, Z_i is a rectangular matrix (with the same dimension as U_i) that accumulates the “momentum” of the iterates U_i . μ is the momentum parameter that weighs how the previous estimates U_i will be mixed with the current estimate U_{i+1} to generate Z_{i+1} . The above iteration is an adaptation of Nesterov’s accelerated first-order method for convex problems [65]. We borrow this momentum formulation, and we study how constant μ selections behave in non-convex problem formulations, such as in (3). *We note that the theory and algorithmic configurations in [65] do not generalize to non-convex problems*, which is one of the contributions of this work.

An important observation is that the factorization $\rho = UU^\dagger$ is not unique. For instance, suppose that U^* is an optimal solution for (3); then, for any rotation matrix $R \in \mathbb{C}^{r \times r}$ satisfying $RR^\dagger = I$, the matrix $\hat{U} = U^*R$ is also optimal for (3).⁵ To resolve this ambiguity, we use the distance between a pair of matrices as the minimum distance $\min_{R \in \mathcal{O}} \|U - U^*R\|_F$ up to rotations, where $\mathcal{O} = \{R \in \mathbb{C}^{r \times r} \mid RR^\dagger = I\}$. In words, we want to track how close an estimate U is to U^* , up to the minimizing rotation matrix.

We are now ready to state the main theorem regarding the MiFGD algorithm:

Theorem 1 (MiFGD convergence rate (Informal)). *Assume that \mathcal{A} satisfies the RIP for some constant $0 < \delta_{2r} < 1$. Let $y = \mathcal{A}(\rho^*)$ denote the set of measurements, by measuring the quantum density matrix ρ^* . Given a good initialization point U_0 and setting step size η and momentum μ appropriately, MiFGD converges with a linear rate to a region—with radius that depends on $O(\mu)$ —around the global solution ρ^* .*

We refer to Theorem 2 of the Methods section for a formal statement. There, we state the conditions under which the simple MiFGD recursion in Eqs. (6)-(7) has a linear convergence rate in iterate distance, up to a constant error level proportional to the momentum parameter μ . The theorem assumes that the observations are noiseless; that is, the observed data is $y = \mathcal{A}(\rho^*)$, where ρ^* is the state of the system. Nevertheless, our experiments suggest that MiFGD is robust to statistical errors and noise in the data. The formal analysis of robustness to noisy data can be derived from our analysis; here, for clarity, we consider this work as the basis for that analysis.

Experimental setup

ρ^* density matrices and quantum circuits⁶

In our numerical and real experiments, we have considered (different subsets of) the following n -qubit pure quantum states:

1. The (generalized) GHZ state:

$$|\text{GHZ}(n)\rangle = \frac{|0\rangle^{\otimes n} + |1\rangle^{\otimes n}}{\sqrt{2}}, \quad n > 2.$$

2. The (generalized) GHZ-minus state:

$$|\text{GHZ}_-(n)\rangle = \frac{|0\rangle^{\otimes n} - |1\rangle^{\otimes n}}{\sqrt{2}}, \quad n > 2.$$

3. The Hadamard state:

$$|\text{Hadamard}(n)\rangle = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}} \right)^{\otimes n}.$$

4. A random state $|\text{Random}(n)\rangle$.

We have implemented these states (on the IBM quantum simulator and/or the IBM’s QPU) using the following circuits. The GHZ state $|\text{GHZ}(n)\rangle$ is generated by applying the Hadamard gate to one of the qubits, and then applying $n - 1$ CNOT gates between this qubit (as a control) and the remaining $n - 1$ qubits (as

⁵To see this, observe that $\rho^* = U^*U^{*\dagger} = U^*IU^{*\dagger} = U^*RR^\dagger U^{*\dagger} = \hat{U}\hat{U}^\dagger$.

⁶The content in this subsection is implemented in the `states.py` component of our complementary software package.

targets). The GHZ-minus state $|\text{GHZ}_-(n)\rangle$ is generated by applying the X gate to one of the qubits (e.g., the first qubit) and the Hadamard gate to the remaining $n - 1$ qubits, followed by applying $n - 1$ CNOT gates between the first qubit (as a target) and the other $n - 1$ qubits (as controls). Finally, we apply the Hadamard gate to all of the qubits. The Hadamard state $|\text{Hadamard}(n)\rangle$ is a separable state, and it is generated by applying the Hadamard gate to all of the qubits. The random state $|\text{Random}(n)\rangle$ is generated by a random quantum gate selection: In particular, for a given circuit depth, we uniformly select among generic single-qubit rotation gates with 3 Euler angles, and controlled- X gates, for every step in the circuit sequence. For the rotation gates, the qubits involved are selected uniformly at random, as well as the angles from the range $[0, 1]$. For the controlled- X gates, the source and target qubits are also selected uniformly at random.

We generically denote the density matrix that correspond to pure state $|\psi\rangle$ as $\rho^\star = |\psi\rangle\langle\psi|$. For clarity, we will drop the bra-ket notation when we refer to $|\text{GHZ}(n)\rangle$, $|\text{GHZ}_-(n)\rangle$, $|\text{Hadamard}(n)\rangle$ and $|\text{Random}(n)\rangle$. While the density matrices of the $\text{GHZ}(n)$ and $\text{GHZ}_-(n)$ are sparse in the $\{|0\rangle, |1\rangle\}^n$ basis, the density matrix of $\text{Hadamard}(n)$ state is fully-dense in this basis, and the sparsity of the density matrix that of $\text{Random}(n)$ may be different from one state to another.

Measuring quantum states⁷

The quantum measurement model. In our experiments (both synthetic and real) we measure the qubits in the Pauli basis [3].⁸ A Pauli basis measurement on an n -qubit system has $d = 2^n$ possible outcomes. The Pauli basis measurement is uniquely defined by the *measurement setting*. A Pauli measurement is a string of n letters $\alpha := (\alpha_1, \alpha_2, \dots, \alpha_n)$ such that $\alpha_k \in \{x, y, z\}$ for all $k \in [n]$. Note that there are at most 3^n distinct Pauli strings. To define the Pauli basis measurement that associated with a given measurement string α , we first define the the following three bases on $\mathbb{C}^{2 \times 2}$:

$$\begin{aligned}\mathcal{B}_x &= \left\{ |x, 0\rangle := \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), |x, 1\rangle := \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \right\}, \\ \mathcal{B}_y &= \left\{ |y, 0\rangle := \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle), |y, 1\rangle := \frac{1}{\sqrt{2}}(|0\rangle - i|1\rangle) \right\}, \\ \mathcal{B}_z &= \{|z, 0\rangle := |0\rangle, |z, 1\rangle := |1\rangle\}.\end{aligned}$$

These are the eigenbases of the single-qubit Pauli operators, σ_x , σ_y , and σ_z , whose 2×2 matrix representation is given by:

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Given a Pauli setting α , the Pauli basis measurement Π_α is defined by the 2^n projectors:

$$\Pi_\alpha = \left\{ |v_\ell^{(\alpha)}\rangle\langle v_\ell^{(\alpha)}| = \bigotimes_{k=1}^n |\alpha_k, \ell_k\rangle\langle \alpha_k, \ell_k| : \ell_k \in \{0, 1\} \forall k \in [1, n] \right\},$$

where ℓ denotes the bit string $(\ell_{k_1}, \ell_{k_2}, \dots, \ell_{k_n})$. Since there are 3^n distinct Pauli measurement settings, there are the same number of possible Pauli basis measurements.

Technically, this set forms a positive operator-valued measure (POVM). The projectors that form Π_α are the measurement outcomes (or POVM elements) and the probability to obtain an outcome $|v_\ell^{(\alpha)}\rangle\langle v_\ell^{(\alpha)}|$ —when the state of the system is ρ^\star —is given by the Born rule: $\langle v_\ell^{(\alpha)} | \rho^\star | v_\ell^{(\alpha)} \rangle = \text{Tr}(|v_\ell^{(\alpha)}\rangle\langle v_\ell^{(\alpha)}| \cdot \rho^\star)$.

The RIP and expectation values of Pauli observables. Starting with the requirements of our algorithm, the sensing mapping $\mathcal{A}(\cdot) : \mathbb{C}^{d \times d} \rightarrow \mathbb{R}^m$ we consider is comprised of a collection of $A_i \in \mathbb{C}^{d \times d}$, $i = 1, \dots, m$ matrices, such that $y_i = \text{Tr}(A_i \rho^\star)$. We denote the vector (y_1, \dots, y_m) by y .

When no prior information about the quantum state is assumed, to ensure its (robust) recovery, one must choose a set m sensing matrices A_i , so that d^2 of them are linearly independent. One example of such choice is the POVM elements of the 3^n Pauli basis measurements.

⁷The content in this subsection is implemented in the `measurements.py` component of our complementary software package.

⁸This is the non-commutative analogue of the Fourier basis, for the case of sparse vectors [78, 17].

Yet, when it is known that the state-to-be-reconstructed is of low-rank, theory on low-rank recovery problems suggests that A_i could just be “incoherent” enough with respect to ρ^* [35], so that recovery is possible from a limited set of measurements, i.e., with $m \ll d^2$. In particular, it is known [60, 35, 36] that if the sensing matrices correspond to random *Pauli monomials*, then $m = O(r \cdot d \cdot \text{poly}(\log d))$. A_i ’s are sufficient for a successful recovery of ρ^* , using convex solvers for (1).⁹ A Pauli monomial P_i is an operator in the set $P_i \in \{\mathbb{1}, \sigma_x, \sigma_y, \sigma_z\}^{\otimes n}$, that is, an n -fold tensor product of single-qubit Pauli operators (including the identity operator). For convenience we relabel the single-qubit Pauli operators as $\sigma_0 := \mathbb{1}, \sigma_1 := \sigma_x, \sigma_2 := \sigma_y$, and $\sigma_3 := \sigma_z$, so that we can also write $P_i = \bigotimes_{k=1}^n \sigma_{i_k}$ with $i_k \in \{0, \dots, 3\}$ for all $k \in [n]$. These results [60, 35, 36] are feasible since the Pauli-monomial-based sensing map $\mathcal{A}(\cdot)$ obeys the RIP property, as in Definition 1.¹⁰ For the rest of the text, we will use the term “*Pauli expectation value*” to denote $\text{Tr}(A_i \rho^*) = \text{Tr}(P_i \rho^*)$.

From Pauli basis measurements to Pauli expectation values. While the theory for compressed sensing was proven for Pauli expectation values, in real QPUs, experimental data is obtained from Pauli basis measurements. Therefore, to make sure we are respecting the compressed sensing requirements on the sensing map, we follow this protocol:

- i) We sample $m = O(r \cdot d \cdot \text{poly}(\log d))$ or $m = \text{measpc} \cdot d^2$ Pauli monomials uniformly over $\{\sigma_i\}^{\otimes n}$ with $i \in \{0, \dots, 3\}$, where $\text{measpc} \in [0, 1]$ represents the percentage of measurements out of full tomography.
- ii) For every monomial, P_i , in the generated set, we identify an experimental setting $\alpha(i)$ that corresponds to the monomial. There, qubits, for which their Pauli operator in P_i is the identity operator, are measured, without loss of generality, in the σ_3 basis. For example, for $n = 3$ and $P_i = \sigma_0 \otimes \sigma_1 \otimes \sigma_1$, we identify the measurement setting $\alpha(i) = (z, x, x)$.
- iii) We measure the quantum state in the Pauli basis that corresponds to $\alpha(i)$, and record the outcomes.

To connect the measurement outcomes to the expectation value of the Pauli monomial, we use the relation:

$$\text{Tr}(P_i \rho^*) = \sum_{\ell \in \{0,1\}^n} (-1)^{\chi_{f(\ell)}} \cdot \text{Tr} \left(|v_\ell^{(\alpha(i))}\rangle \langle v_\ell^{(\alpha(i))}| \cdot \rho^* \right), \quad (8)$$

where $f(\ell) : \{0, 1\}^n \rightarrow \{0, 1\}^n$ is a mapping that takes a bit string ℓ and returns a new bit string $\tilde{\ell}$ (of the same size) such that $\tilde{\ell}_k = 0$ for all k ’s for which $i_k = 0$ (that is, the locations of the identity operators in P_i), and $\chi_{\tilde{\ell}}$ is the parity of the bit string $\tilde{\ell}$.

Algorithmic setup

In our implementation, we expose a number of control parameters, including the maximum number of iterations `maxiters`, the learning rate η , the relative error from successive state iterates `reltol`, the acceleration parameter μ , the percentage of the complete set of measurements (i.e. over all possible Pauli monomials) `measpc`, and the seed. In the sequel experiments we set `maxiters` = 1000, $\eta = 10^{-3}$, `reltol` = 5×10^{-4} unless stated differently. Regarding acceleration, $\mu = 0$ when acceleration is muted; we experiment over the range of values $\mu \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{3}, \frac{3}{4}\}$ when investigating the acceleration effect, beyond the theoretically suggested μ^* . In order to explore the dependence of our approach on the number of measurements available, `measpc` varies over the set of {5%, 10%, 15%, 20%, 40%, 60%}; `seed` is used for differentiating repeating runs with all other parameters kept fixed.¹¹

Denoting $\hat{\rho}$ the estimate of ρ^* by `MIFGD`, we report on outputs including:

- The evolution with respect to the distance of current state to the target one: $\|\hat{\rho} - \rho^*\|_F$, for various μ ’s.
- The number of iterations to reach `reltol` for various μ ’s and for different circuits’ target states (ρ^*).

⁹The main difference between [35, 36] and [60] is that the former guarantees recovery for almost all choices of $m = O(r \cdot d \cdot \text{poly}(\log d))$ random Pauli monomials, while the latter proves that there exists a *universal* set of $m = O(r \cdot d \cdot \text{poly}(\log d))$ Pauli monomials A_i that guarantees successful recovery.

¹⁰In particular, the RIP is satisfied for the sensing mechanisms that obeys $(\mathcal{A}(\rho^*))_i = \frac{d}{\sqrt{m}} \text{Tr}(A_i^* \rho^*)$, $i = 1, \dots, m$. Further, the case considered in [60] holds for a slightly larger set than the set of rank- r density matrices: for all $\rho \in \mathbb{C}^{d \times d}$ such that $\|\rho\|_* \leq \sqrt{r} \|\rho\|_F$.

¹¹`maxiters` is `num.iterations` in the code; also `reltol` is `relative.error.tolerance`, `measpc` is `complete.measurements.percentage`.

- The fidelity of $\hat{\rho}$, defined as $\text{Tr}(\rho^* \hat{\rho})$ (for rank-1 ρ^*), as a function of the acceleration parameter μ in the default set and for different circuits' target states (ρ^*).

In our plots, we sweep over our default set of `measpc` values, repeat 5 times for each individual setup, varying supplied seed, and depict their 25-, 50- and 75-percentiles.

Experimental setup on quantum processing unit (QPU)

We show empirical results on 6- and 8-qubit real data, obtained on the 20-qubit IBM QPU `ibmq_boeblingen`. The layout/connectivity of the device is shown in Figure 1. The 6-qubit data was from qubits `[0, 1, 2, 3, 8, 9]`, and the 8-qubit data was from `[0, 1, 2, 3, 8, 9, 6, 4]`. The T_1 coherence times are `[39.1, 75.7, 66.7, 100.0, 120.3, 39.2, 70.7, 132.3]` μs , and T_2 coherence times are `[86.8, 94.8, 106.8, 63.6, 156.5, 66.7, 104.5, 134.8]` μs . The circuit for generating 6-qubit and 8-qubit GHZ states are shown in Fig 1. The typical two qubit gate errors measured from randomized benchmarking (RB) for relevant qubits are summarized in Table 1.

$C_0 X_1$	$C_1 X_2$	$C_2 X_3$	$C_3 X_8$	$C_8 X_9$	$C_3 X_4$	$C_1 X_6$
0.0072	0.0062	0.0087	0.0077	0.0152	0.0167	0.0133

Table 1: Two qubit error rates for the relevant gates used in generating 6-qubit and 8-qubit GHZ states on `ibmq_boeblingen`.

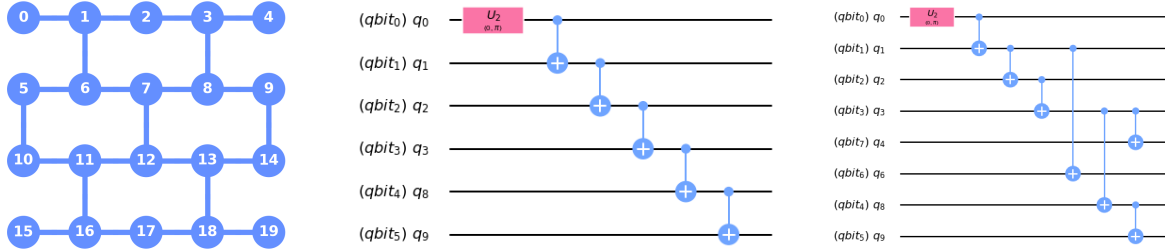


Figure 1: **Left panel:** Layout connectivity of IBM backend `ibmq_boeblingen`; **Middle and right panels:** Circuits used to generate 6-qubit state (left) and 8-qubit GHZ state (right). *qbit* refers to the quantum registers used in qiskit, and *q* corresponds to qubits on the real device.

The QST circuits were generated using the tomography module in `qiskit-ignis`.¹² For complete QST of a n -qubits state 2^n circuits are needed. The result of each circuit is averaged over 8192, 4096 or 2048, for different n -qubit scenarios. To mitigate for readout errors, we prepare and measure all of the 2^n computational basis states in the computation basis to construct a calibration matrix C . C has dimension 2^n by 2^n , where each column vector corresponds to the measured outcome of a prepared basis state. In the ideal case of no readout error, C is an identity matrix. We use C to correct for the measured outcomes of the experiment by minimizing the function:

$$\min_{v^{\text{cal}} \in \mathbb{R}^d} \|C v^{\text{cal}} - v^{\text{meas}}\|^2 \quad \text{subject to} \quad \sum_i v_i^{\text{cal}} = 1, v_i^{\text{cal}} \geq 0, \forall i = 1, \dots, d \quad (9)$$

Here v^{meas} and v^{cal} are the measured and calibrated outputs, respectively. The minimization problem is then formulated as a convex optimization problem and solved by quadratic programming using the package `cvxopt` [89].

MiFGD on 6- and 8-qubit real quantum data

We realize two types of quantum states on IBM QPUs, parameterized by the number of qubits n for each case: these are the `GHZ_(n)` and `Hadamard(n)` circuits. We collected measurements over all possible Pauli settings by repeating the experiment for each setting a number of times: these are the number of `shots` for each setting. The (circuit, number of `shots`) measurement configurations from IBM Quantum devices are summarized in Table 2.

¹²<https://github.com/Qiskit/qiskit-ignis>.

In the Appendix, we provide target error list plots for the evolution of $\|\hat{\rho} - \rho^*\|_F^2$ for reconstructing all the settings in Table 2, both for real data and for simulated scenarios. Further, we provide plots that relate the effect of momentum acceleration on the final fidelity observed for these cases. For clarity, in Figure 2, we summarize the efficiency of momentum acceleration, by showing the reconstruction error only for the following settings: $\text{maxiters} = 1000$, $\eta = 10^{-3}$, $\text{reltol} = 5 \times 10^{-4}$, and $\text{measpc} = 20\%$. In the plots, $\mu = 0$ corresponds to the FGD algorithm in [73], μ^* corresponds to the value obtained through our theory, while we use $\mu \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{3}, \frac{3}{4}\}$ to study the acceleration effect. For μ^* , per our theory, we follow the rule for μ ; see also the Methods section for more details.¹³ We run each QST experiment for 5 times for random initializations. We record the evolution of the $\|\hat{\rho} - \rho^*\|_F^2$ error at each step, and stop when the relative error of successive iterates gets smaller than reltol or the number of iterations exceeds maxiters (whichever happens first). To implement $\text{measpc} = 20\%$, we follow the description given above Eq. (8) with $m = \text{measpc} \cdot d^2$.

To highlight the level of noise existing in real quantum data, in Figure 3, we repeat the same setting using the QASM simulator in `qiskit-aer`. This is a parallel, high performance quantum circuit simulator written in C++ that can support a variety of realistic circuit level noise models.

Figure 2 summarizes the performance of our proposal on different ρ^* , and for different μ values on real IBM QPU data. All plots show the evolution of $\|\hat{\rho} - \rho^*\|_F$ across iterations, featuring a steep dive to convergence for the largest value of μ we tested: we report that we also tested $\mu = 0$, which shows only slight worse performances than μ^* . Figure 2 highlights the universality of our approach: its performance is oblivious to the quantum state reconstructed, as long as it satisfies purity or it is close to a pure state. Our method does not require any additional structure assumptions in the quantum state.

To highlight the effect of real noise on the performance of `MiFGD`, we further plot its performance on the same settings but using measurements coming from an idealized quantum simulator. Figure 3 considers the exact same settings as in Figure 2. It is obvious that `MiFGD` can achieve better reconstruction performance when data are less erroneous. This also highlights that, in real noisy scenarios, the radius of the convergence region of `MiFGD` around ρ^* is controlled mostly by the noise level, rather than by the inclusion of momentum acceleration.

Finally, in Figure 4, we depict the fidelity of $\hat{\rho}$ achieved using `MiFGD`, defined as $\text{Tr}(\rho^* \hat{\rho})$, versus various μ values and for different circuits (ρ^*). Shaded area denotes standard deviation around the mean over repeated runs in all cases. The plots show the significant gap in performance when using real quantum data versus using synthetic simulated data within a controlled environment.

Circuit	# shots
GHZ ₋ (6)	2048
GHZ ₋ (6)	8192
GHZ ₋ (8)	2048
GHZ ₋ (8)	4096
Hadamard(6)	8192
Hadamard(8)	4096

Table 2: QPU settings.

Performance comparison with tomography methods in Qiskit

We compare `MiFGD` with publicly available implementations for QST reconstruction. Two common techniques for QST, included in the `qiskit-ignis` distribution [75], are: *i*) the `CVXPY` fitter method, that uses the `CVXPY` convex optimization package [24, 5]; and *ii*) the `lstsq` method, that uses least-squares fitting [81]. Both methods solve the *full tomography problem*¹⁴ according to the following expression:

$$\begin{aligned} \min_{\rho \in \mathbb{C}^{d \times d}} \quad & f(\rho) := \frac{1}{2} \|\mathcal{A}(\rho) - y\|_2^2 \\ \text{subject to} \quad & \rho \succeq 0, \text{Tr}(\rho) = 1. \end{aligned} \tag{10}$$

We note that `MiFGD` is not restricted to “tall” U scenarios to encode PSD and rank constraints: even without rank constraints, one could still exploit the matrix decomposition $\rho = UU^\dagger$ to avoid the PSD projection, $\rho \succeq 0$, where $U \in \mathbb{C}^{d \times d}$. For the `lstsq` fitter method, the putative estimate $\hat{\rho}$ is rescaled using the method

¹³For this application, $\sigma_r(\rho^*) = 1$, $\tau(\rho^*) = 1$, and $r = 1$ by construction; we also approximated $\kappa = 1.223$, which, for user-defined $\varepsilon = 1$, results in $\mu^* = 4.5 \cdot 10^{-4}$. Note that smaller ε values result into a smaller radius of the convergence region; however, more pessimistic ε values result into small μ , with no practical effect in accelerating the algorithm.

¹⁴In Ref. [46] it was shown that the minimization program (10) yields a robust estimation of low-rank states in the compressed sensing. Thus, one can use `CVXPY` fitter method to solve (10) with $m \ll d^2$ Pauli expectation value to obtain a robust reconstruction of ρ^* .

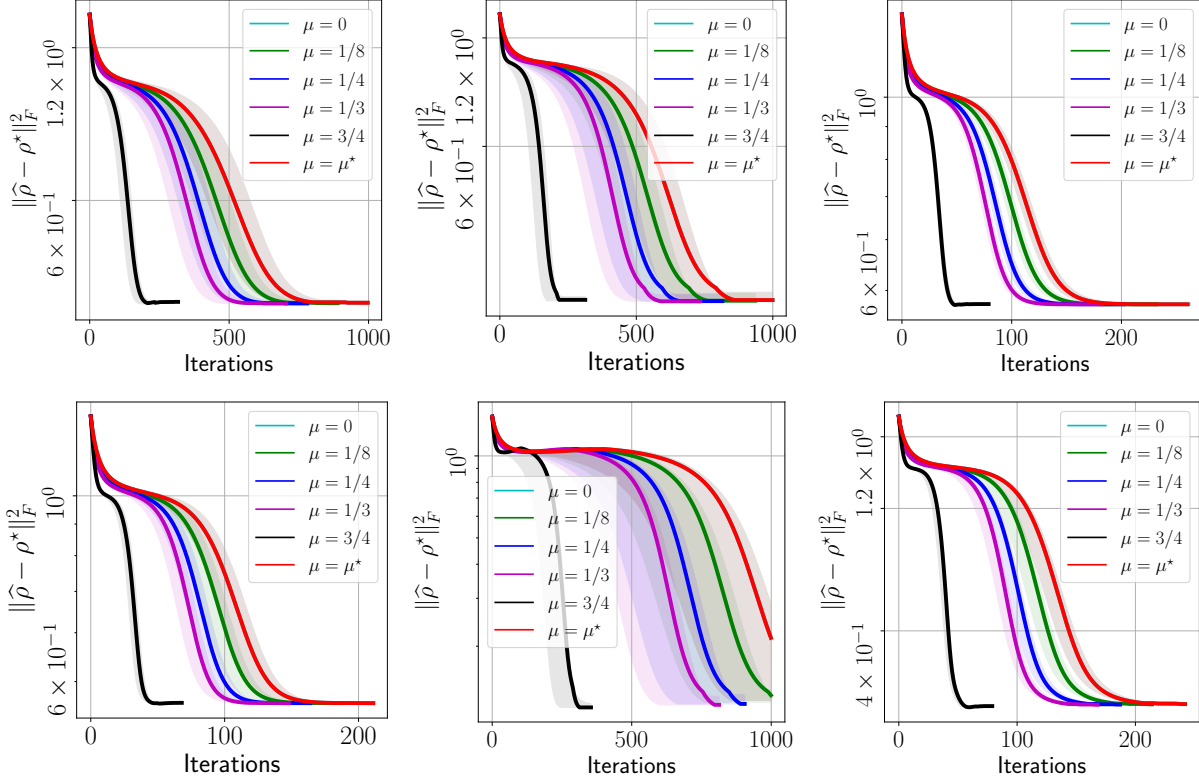


Figure 2: Target error list plots $\|\hat{\rho} - \rho^*\|_F^2$ versus method iterations using real IBM QPU data. **Top-left:** GHZ₋₍₆₎ with 2048 shots; **Top-middle:** GHZ₋₍₆₎ with 8192 shots; **Top-right:** GHZ₋₍₈₎ with 2048 shots; **Bottom-left:** GHZ₋₍₈₎ with 4096 shots/copies of ρ^* ; **Bottom-middle:** Hadamard₋₍₆₎ with 8192 shots; **Bottom-right:** Hadamard₋₍₈₎ with 4096 shots. All cases have $\text{measpc} = 20\%$. Shaded area denotes standard deviation around the mean over repeated runs in all cases.

proposed in [81]. For CVXPY, the convex constraint makes the optimization problem a semidefinite programming (SDP) instance. By default, CVXPY calls the SCS solver that can handle all problems (including SDPs) [68, 69]. Further comparison results with matrix factorization techniques from the machine learning community is provided in the Appendix for $n = 12$.

The settings we consider for full tomography are the following: GHZ(n), Hadamard(n) and Random(n) quantum states (for $n = 3, \dots, 8$). We focus on fidelity of reconstruction and computation timings performance between CVXPY, lstsq and MiFGD. We use 100% of the measurements. We experimented with states simulated in QASM and measured taking 2,048 shots. For MiFGD, we set $\eta = 0.001$, $\mu = \frac{3}{4}$, and stopping criterion/tolerance $\text{reltol} = 10^{-5}$. All experiments are run on a Macbook Pro with 2.3 GHz Quad-Core Intel Core i7CPU and 32GB RAM.

The results are shown in Figure 5; higher-dimensional cases are provided in Table 3. Some notable remarks: *i*) For small-scale scenarios ($n = 3, 4$), CVXPY and lstsq attain almost perfect fidelity, while being comparable or faster than MiFGD. *ii*) The difference in performance becomes apparent from $n = 6$ and on: while MiFGD attains 98% fidelity in < 5 seconds, CVXPY and lstsq require up to hundreds of seconds to find a good solution. *iii*) Finally, while MiFGD gets to high-fidelity solutions in seconds for $n = 7, 8$, CVXPY and lstsq methods require more than 12 hours execution time; however, their execution never ended, since the memory usage exceeded the system's available memory.

It is noteworthy that the reported fidelities for MiFGD are the fidelities at the last iteration, before the stopping criterion is activated, or the maximum number of iterations is exceeded. However, the reported fidelity is not necessarily the best one during the whole execution: for all cases, we observe that MiFGD

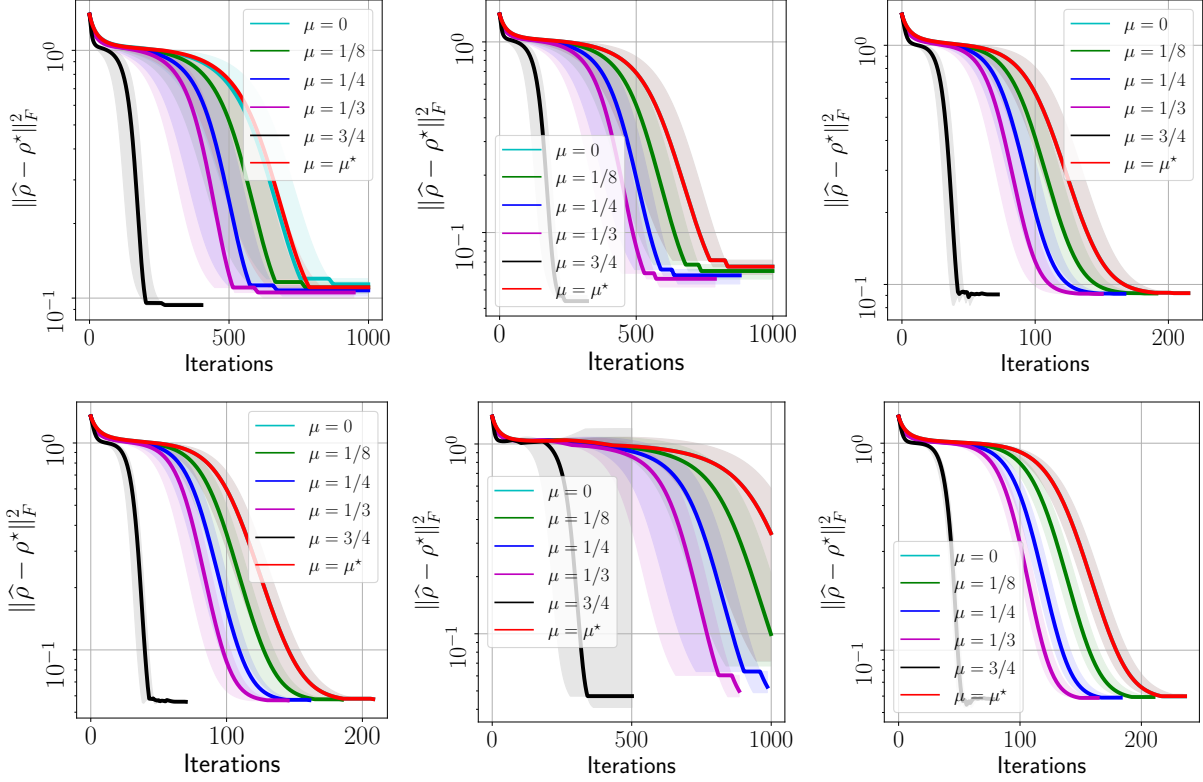


Figure 3: Target error list plots $\|\hat{\rho} - \rho^*\|_F^2$ versus method iteration using synthetic IBM's quantum simulator data. **Top-left:** GHZ₋₍₆₎ with 2048 shots; **Top-middle:** GHZ₋₍₆₎ with 8192 shots; **Top-right:** GHZ₋₍₈₎ with 2048 shots; **Bottom-left:** GHZ₋₍₈₎ with 4096 shots; **Bottom-middle:** Hadamard(6) with 8192 shots; **Bottom-right:** Hadamard(8) with 4096 shots. All cases have $\text{measpc} = 20\%$. Shaded area denotes standard deviation around the mean over repeated runs in all cases.

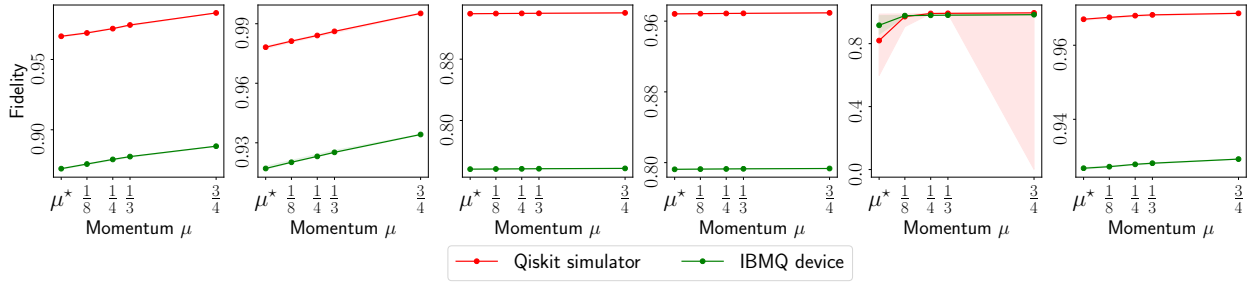


Figure 4: Fidelity list plots where we depict the fidelity of $\hat{\rho}$ to ρ^* . From left to right: *i*) GHZ₋₍₆₎ with 2048 shots; *ii*) GHZ₋₍₆₎ with 8192 shots; *iii*) GHZ₋₍₈₎ with 2048 shots; *iv*) GHZ₋₍₈₎ with 4096 shots; *v*) Hadamard(6) with 8192 shots; *vi*) Hadamard(8) with 4096 shots. All cases have $\text{measpc} = 20\%$. Shaded area denotes standard deviation around the mean over repeated runs in all cases.

finds intermediate solutions with fidelity $> 99\%$. Though, it is not realistic to assume that the iteration with the best fidelity is known a priori, and this is the reason we report only the final iteration fidelity.

Circuit	Method	Fidelity	Time (secs)
GHZ(7)	MiFGD	0.969397	10.6709
Hadamard(7)	MiFGD	0.969397	10.4926
Random(7)	MiFGD	0.968553	9.59607
All above	lstsq, CVXPY	> 12h + Memory limit exceed	
GHZ(8)	MiFGD	0.940389	35.0666
Hadamard(8)	MiFGD	0.940390	37.5331
Random(8)	MiFGD	0.942815	36.3251
All above	lstsq, CVXPY	> 12h + Memory limit exceed	

Table 3: Fidelity of reconstruction and computation timings using 100% of the complete measurements. Rows correspond to combinations of number of qubits (3 ~ 8), synthetic circuit, and tomographic method (MiFGD, Qiskit’s lstsq and CVXPY fitters. 2048 shots per measurement circuit. For MiFGD, $\eta = 0.001$, $\mu = \frac{3}{4}$, $\text{reltol} = 10^{-5}$. “-” indicates the procedure did not finish running in 48 hours. All experiments are run on a 13” Macbook Pro with 2.3 GHz Quad-Core Intel Core i7 CPU and 32 GB RAM.

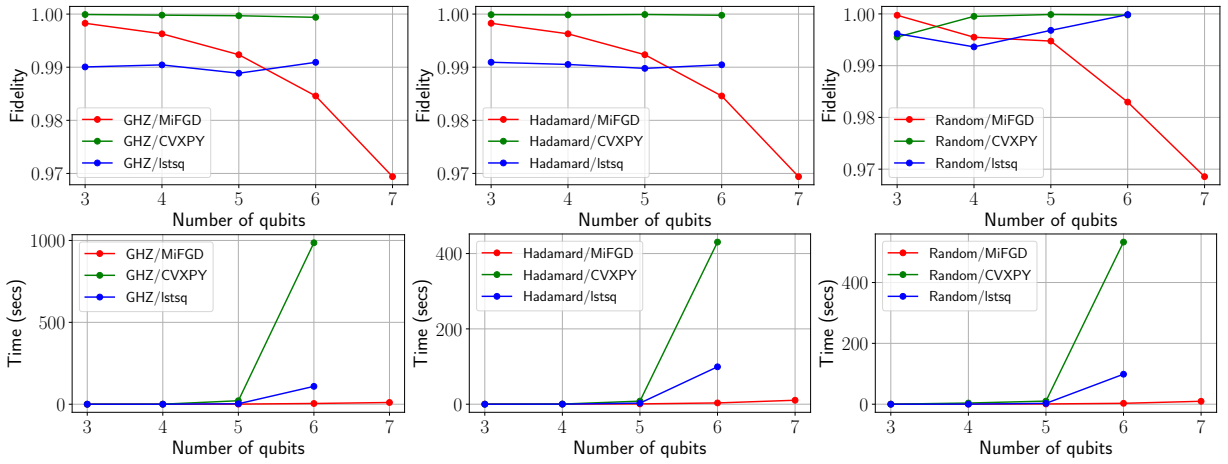


Figure 5: Fidelity versus time plots using synthetic IBM’s quantum simulator data. **Left panel:** GHZ_(n) for $n = 3, 4$; **Middle panel:** Hadamard_(n) for $n = 3, 4$; **Right panel:** Random_(n) for $n = 3, 4$.

Performance comparison of MiFGD with neural-network quantum state tomography

We compare the performance of MiFGD with neural network approaches. Per [86, 10, 87, 29], we model a quantum state with a two-layer Restricted Boltzmann Machine (RBM). RBMs are stochastic neural networks, where each layer contains a number of binary stochastic variables: the size of the visible layer corresponds to the number of input qubits, while the size of the hidden layer is a hyperparameter controlling the representation error. We experiment with three types of RBMs for reconstructing either the positive-real wave function, the complex wave function, or the density matrix of the quantum state. In the first two cases the state is assumed pure while in the last, general mixed quantum states can be represented. We leverage the implementation in QuCumber [10], PositiveRealWaveFunction (PRWF), ComplexWaveFunction (CWF), and DensityMatrix (DM), respectively.

We reconstruct GHZ(n), Hadamard(n) and Random(n) quantum states (for $n = 3, \dots, 8$), by training PRWF, CWF, and DM neural networks¹⁵ with measurements collected by the QASM Simulator.

For our setting, we consider $\text{measpc} = 50\%$ and $\text{shots} = 2048$. The set of measurements is presented to the RBM implementation, along with the target positive-real wave function (for PRWF), complex wave-function (for CWF) or the target density matrix (for DM) in a suitable format for training. We train Hadamard and Random states with 20 epochs, and GHZ state with 100 epochs.¹⁶ We set the number of hidden variables

¹⁵We utilize GPU (NVidia GeForce GTX 1080 TI, 11GB RAM) for faster training of the neural networks.

¹⁶We experimented higher number of epochs (up to 500) for all cases, but after the reported number of epochs, QuCumber methods

(and also of additional auxiliary variables for DM) to be equal to the number of input variables n and we use 100 data points for both the positive and the negative phase of the gradient (as per the recommendation for the defaults). We choose $k = 10$ contrastive divergence steps and fixed the learning rate to 10 (per hyperparameter tuning). Lastly, we limit the fitting time of Qucumber methods (excluding data preparation time) to be three hours. To compare to the RBM results, we run MiFGD with $\eta = 0.001$, $\mu = \frac{3}{4}$, $\text{reltol} = 10^{-5}$ and using $\text{measpc} = 50\%$, keeping previously chosen values for all other hyperparameters.

We report the fidelity of the reconstruction as a function of elapsed training time for $n = 3, 4$ in Figure 6 for PRWF, CWF, and DM. We observe that for all cases, Qucumber methods are orders of magnitude slower than MiFGD. E.g., for $n = 8$, for all three states, CWF and DM did not finish a single epoch in 3 hours, while MiFG achieves high fidelity in less than 30 seconds. For the Hadamard(n) and Random(n), reaching reasonable fidelities is significantly slower for both CWF and DM, while PRWF hardly improves its performance throughout the training. For the GHZ case, CWF and DM also shows *non-monotonic* behaviors: even after a few *thousands of seconds*, fidelities have not “stabilized”, while PRWF stabilizes in very low fidelities. In comparison MiFGD is several orders of magnitude faster than both CWF and DM and fidelity smoothly increases to comparable or higher values.

What is notable is the scalability of MiFGD compared to neural network approaches for higher values of n . In Table 4, we report *final* fidelities (within the 3 hour time window), and reported times. For many cases, CWF and DM methods did not complete a single iterations within 3 hours. In contrast, MiFGD obtained high-fidelity solutions within seconds.

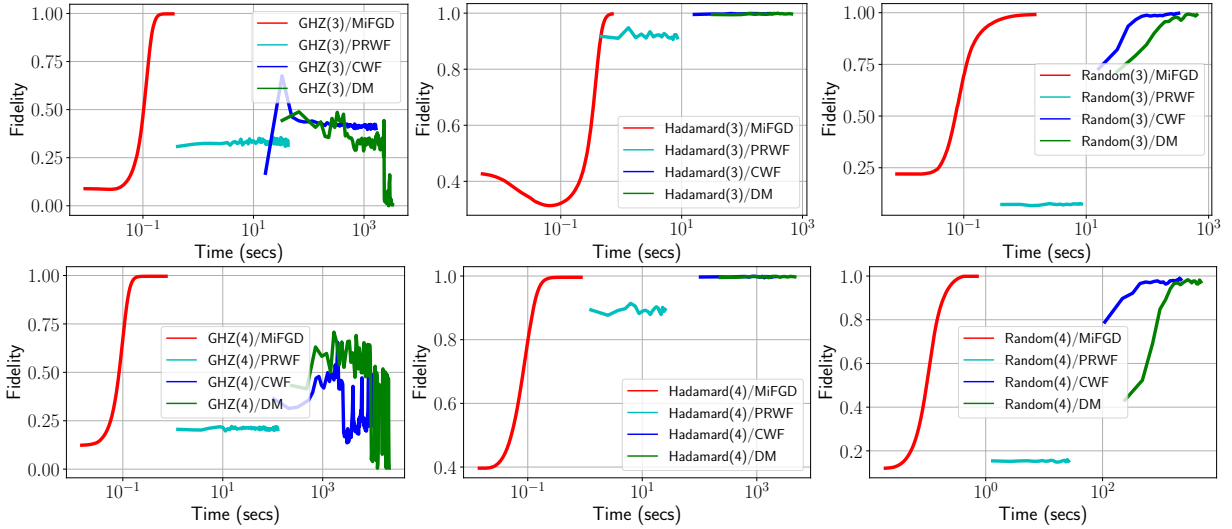


Figure 6: Fidelity versus time plots on MiFGD, PRWF, CWF, and DM, using synthetic IBM’s quantum simulator data. **Left panel:** GHZ(n) for $n = 3, 4$; **Middle panel:** Hadamard(n) for $n = 3, 4$; **Right panel:** Random(n) for $n = 3, 4$.

The effect of parallelization

We study the effect of parallelization in running MiFGD. We parallelize the iteration step across a number of processes, that can be either distributed and network connected, or sharing memory in a multicore environment. Our approach is based on Message Passing Interface (MPI) specification [28], which is the lingua franca for interprocess communication in high performance parallel and supercomputing applications. A MPI implementation provides facilities for launching processes organized in a virtual topology and highly tuned primitives for point-to-point and collective communication between them.

We assign to each process a subset of the measurement labels consumed by the parallel computation. At each step, a process first computes the local gradient-based corrections due only to its assigned Pauli

did not improve, if not worsened.

Circuit		Method			
		MiFGD	PRWF	CWF	DM
GHZ(3)	Fidelity	0.997922	0.314167	0.401737	0.005389
	Time (secs)	0.348652	42.27607	1649.224	3279.118
Hadamard(3)	Fidelity	0.997229	0.912268	0.997914	0.997222
	Time (secs)	0.706872	8.492405	325.7040	656.6696
Random(3)	Fidelity	0.991063	0.074774	0.997493	0.989754
	Time (secs)	1.447057	8.345135	322.4730	640.8185
GHZ(4)	Fidelity	0.996029	0.204313	0.276491	0.138459
	Time (secs)	0.733128	126.2749	10756.87	> 3h
Hadamard(4)	Fidelity	0.996078	0.894883	0.998071	0.997389
	Time (secs)	0.852895	25.15520	2087.540	4613.964
Random(4)	Fidelity	0.998850	0.152971	0.984164	0.972877
	Time (secs)	0.713302	26.18863	2185.091	4802.495
GHZ(5)	Fidelity	0.992105	0.132725	0.274665	0.005138
	Time (secs)	0.946350	395.3379	> 3h	> 3h
Hadamard(5)	Fidelity	0.992102	0.869603	0.998246	0.996516
	Time (secs)	1.183290	79.39444	9319.140	> 3h
Random(5)	Fidelity	0.995126	0.015913	0.623273	0.086777
	Time (secs)	0.988173	79.22450	9275.836	> 3h
GHZ(6)	Fidelity	0.984352	0.089355	0.437323	0.310067
	Time (secs)	3.829866	1167.985	> 3h	> 3h
Hadamard(6)	Fidelity	0.984384	0.842515	0.990849	0.998077
	Time (secs)	2.500354	246.0011	> 3h	> 3h
Random(6)	Fidelity	0.989543	0.143145	0.784873	0.302534
	Time (secs)	1.991154	237.7037	> 3h	> 3h
GHZ(7)	Fidelity	0.969174	0.058387	0.080648	N/A
	Time (secs)	6.174129	3633.082	> 3h	> 3h
Hadamard(7)	Fidelity	0.969156	0.818174	0.996586	N/A
	Time (secs)	6.324469	713.9404	> 3h	> 3h
Random(7)	Fidelity	0.967640	0.141745	0.06568	N/A
	Time (secs)	6.802577	746.2630	> 3h	> 3h
GHZ(8)	Fidelity	0.940601	0.0400391	N/A	N/A
	Time (secs)	21.16011	> 3h	> 3h	> 3h
Hadamard(8)	Fidelity	0.940638	0.794892	N/A	N/A
	Time (secs)	22.30246	2344.796	> 3h	> 3h
Random(8)	Fidelity	0.939418	0.050521	N/A	N/A
	Time (secs)	22.81059	2196.259	> 3h	> 3h

Table 4: Fidelity of reconstruction and computation timings using $\text{measpc}=50\%$ and $\text{shots}=2048$. Rows correspond to combinations of number of qubits ($3 \sim 8$), final fidelity within the 3h time limit, and computation time. For MiFGD, $\eta = 0.001$, $\mu = \frac{3}{4}$, $\text{tol} = 10^{-5}$. “N/A” indicates that the method could not complete a single epoch in 3 hour training time limit, and thus could not provide any fidelity result. All experiments are run on a NVidia GeForce GTX 1080 TI, 11GB RAM.

monomials and corresponding measurements. These local gradient-based corrections will then (i) need to be communicated, so that they can be added, and (ii) finally, their sum will be shared across all processes to produce a global update for U for next step. We accomplish this structure in MPI using `MPI_Allreduce` collective communication primitive with `MPI_SUM` as its reduction operator: the underlying implementation will ensure minimum communication complexity for the operation (e.g. $\log p$ steps for p processes organized in a communication ring) and thus maximum performance.¹⁷ We leverage `mpi4py` [23] bindings to issue MPI calls in our parallel Python code.

We conducted our parallel experiments on a server equipped with 4 x E7-4850 v2 CPUs @ 2.30GHz (48/96 physical/virtual cores), 256 GB RAM, using shared memory multiprocessing over multiple cores. We experimented with states simulated in QASM and measured taking 8,192 shots; parallel MiFGD runs with default parameters and using all measurements (`measpc=100%`). Reported times are wall-clock computation time. These exclude initialization time for all processes to load Pauli monomials and measurements: we here target parallelizing computation proper in MiFGD.

In our first round of experiments, we investigate the scalability of our approach. We vary the number p of parallel processes ($p = 1, 2, 4, 8, 16, 32, 48, 64, 80, 96$), collect timings for reconstructing `GHZ(4)`, `Random(6)` and `GHZMinus(8)` states and report speedups T_p/T_1 we gain from MiFGD in Figure 7(Left panel). We observe that the benefits of parallelization are pronounced for bigger problems (here: $n = 8$ qubits) and maximum scalability results when we use all physical cores (48 in our platform).

Further, we move to larger problems ($n = 10$ qubits: reporting on reconstructing `Hadamard(10)` state) and focus on the effect parallelization to achieving a given level of fidelity in reconstruction. In Figure 7(Middle panel) we illustrate the fidelity as a function of the time spent in the iteration loop of MiFGD for ($p = 8, 16, 32, 48, 64$): we observe the smooth path to convergence in all p counts which again minimizes compute time for $p = 48$. Note that in this case we use `measpc = 10%` and $\mu = \frac{1}{4}$.

Finally, in Figure 7(Right panel), we fix the number of processes to $p = 48$, in order to minimize compute time and increase the percentage of used measurements to 20% of the total available for `Hadamard(10)`. We vary the acceleration parameter, $\mu = 0$ (no acceleration) to $\mu = \frac{1}{4}$ and confirm that we indeed get faster convergence times in the latter case while the fidelity value remains the same (i.e. coinciding upper plateau value in the plots). We can also compare with the previous fidelity versus time plot, where the same μ but half the measurements are consumed: more measurements translate to faster convergence times (plateau is reached roughly 25% faster; compare the green line with the yellow line in the previous plot).

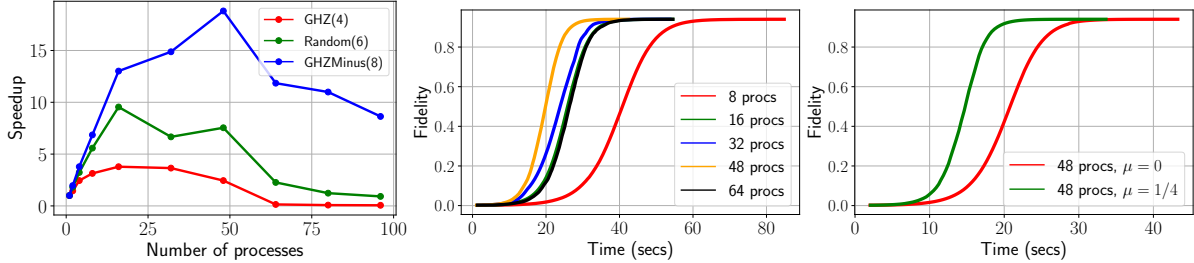


Figure 7: **Left panel:** Scalability of our approach as we vary the number p of parallel processes. **Middle panel:** Fidelity function versus time consumed for different number of processes p . **Right panel:** The effect of momentum for a fixed scenario with `Hadamard(10)` state, $p = 48$, and varying momentum from $\mu = 0$ to $\mu = \frac{1}{4}$.

¹⁷This communication pattern can alternatively be realized in two stages, as naturally suggested in its structure: (i) first invoke MPI's `MPI_Reduce` primitive, with `MPI_SUM` as its reduction operator, which results in the element-wise accumulation of local corrections (vector sum) at a single, designated *root* process, and (ii) finally, send a "copy" of this sum from *root* process to each process participating in the parallel computation (broadcasting); `MPI_Bcast` primitive can be utilized for this latter stage. However, `MPI_Allreduce` is typically faster, since its actual implementation is not constrained by the requirement to have the sum available at a specific, *root* process, at an intermediate time point - as the two-stage approach implies.

Discussion

We have introduced the `MiFGD` algorithm for the factorized form of the low-rank QST problems. We proved that, under certain assumptions on the problem parameters, `MiFGD` converges linearly to a neighborhood of the optimal solution, whose size depends on the momentum parameter μ , while using acceleration motions in a non-convex setting. We demonstrate empirically, using both simulated and real data, that `MiFGD` outperforms non-accelerated methods on both the original problem domain and the factorized space, contributing to recent efforts on testing QST algorithms in real quantum data [77]. These results expand on existing work in the literature illustrating the promise of factorized methods for certain low-rank matrix problems. Finally, we provide a publicly available implementation of our approach, compatible to the open-source software Qiskit [75], where we further exploit parallel computations in `MiFGD` by extending its implementation to enable efficient, parallel execution over shared and distributed memory systems.

Despite our theory does not apply to the Pauli basis measurement directly (i.e., *using randomly selected Pauli bases Π_α , does not lead to the ℓ_2 -norm RIP*), using the data from random Pauli basis measurements directly could provide excellent tomographic reconstruction with `MiFGD`. Preliminary results suggest that only $O(r \cdot \log d)$ random Pauli bases should be taken for a reconstruction, with the same level of accuracy as with $O(r \cdot d \cdot \log d)$ expectation values of random Pauli matrices. We leave the analysis of our algorithm in this case for future work, along with detailed experiments.

Related Work

Matrix sensing. The problem of low-rank matrix reconstruction from few samples was first studied within the paradigm of convex optimization, using the nuclear norm minimization [76, 57, 61]. The use of non-convex approaches for low-rank matrix recovery—by imposing rank-constraints—have been proposed in [44, 58, 53]. In all these works, the convex and non-convex algorithms involve a full, or at least a truncated, singular value decomposition (SVD) per algorithm iteration. Since SVD can be prohibitive, these methods are limited to relatively small system sizes.

Momentum acceleration methods are used regularly in the convex setting, as well as in machine learning practical scenarios [50, 85, 50, 11, 67, 15, 32]. While momentum acceleration was previously studied in non-convex programming setups, it mostly involve non-convex constraints with a convex objective function [52, 53, 49, 94]; and generic non-convex settings but only considering with the question of whether momentum acceleration leads to fast convergence to a saddle point or to a local minimum, rather than to a global optimum [31, 56, 18, 4].

The factorized version for semi-definite programming was popularized in [16]. Effectively the factorization of a the set of PSD matrices to a product of rectangular matrices results in a non-convex setting. This approach have been heavily studied recently, due to computational and space complexity advantages [43, 19, 95, 96, 88, 72, 74, 82, 12, 13, 73, 30, 40, 54]. None of the works above consider the inclusion and analysis of momentum. Moreover, the *Procrustes Flow* approach [96, 88] uses certain initializations techniques, and thus relies on multiple SVD computations. Our approach on the other hand uses a single, unique, top- r SVD computation. Comparison results beyond QST are provided in the appendix.

Compressed sensing QST using non-convex optimization. There are only few works that study non-convex optimization in the context of compressed sensing QST. The authors of [79] propose a hybrid algorithm that *i*) starts with a conjugate-gradient (CG) algorithm in the factored space, in order to get initial rapid descent, and *ii*) switch over to accelerated first-order methods in the original ρ space, provided one can determine the switch-over point cheaply. Using the multinomial maximum likelihood objective, in the initial CG phase, the Hessian of the objective is computed per iteration (i.e., a $4^n \times 4^n$ matrix), along with its eigenvalue decomposition. Such an operation is costly, even for moderate values of qubit number n , and heuristics are proposed for its completion. From a theoretical perspective, [79] provide no convergence or convergence rate guarantees.

From a different perspective, [66] relies on spectrum estimation techniques [37, 20] and the Empirical Young Diagram algorithm [6, 48] to prove that $O(rd/\varepsilon)$ copies suffice to obtain an estimate $\hat{\rho}$ that satisfies $\|\hat{\rho} - \rho^*\|_F^2 \leq \varepsilon$; however, to the best of our knowledge, there is no concrete implementation of this technique to compare with respect to scalability.

Ref. [77] presented an experimental implementation of compressed sensing QST of a $n = 7$ qubit system, where only 127 Pauli basis measurements are available. To achieve recovery in practice, the authors proposed a computationally efficient estimator, based on gradient descent method in the factorized space. The authors of [77] focus on the experimental efficiency of the method, and provide no specific results on the optimization efficiency, neither convergence guarantees of the algorithm. Further, there is no available implementation publicly available.

Similar to [77], Ref. [54] also proposes a non-convex projected gradient decent algorithm that works on the factorized space in the QST setting. The authors prove a rigorous convergence analysis and show that, under proper initialization and step-size, the algorithm is guaranteed to converge to the global minimum of the problem, thus ensuring a provable tomography procedure. *Our results extend these results by including acceleration techniques in the factorized space.* The key contribution of our work is proving convergence of the proposed algorithm in a *linear* rate to the global minimum of the problem, under common assumptions. Proving our results required developing a whole set of new techniques, which are not based on a mere extension of existing results.

Compressed sensing QST using convex optimization. The original formulation of compressed sensing QST [36] is based on convex optimization methods, solving the trace-norm minimization, to obtain an estimation of the low-rank state. It was later shown [46] that essentially any convex optimization program can be used to robustly estimate the state. In general, there are two drawbacks in using convex optimization optimization in QST. Firstly, as the dimension of density matrices grow exponentially in the number of qubits, the search space in convex optimization grows exponentially in the number of qubits. Secondly, the optimization requires projection onto the PSD cone at every iteration, which becomes exponentially hard in the number of qubits. *We avoid these two drawbacks by working in the factorized space.* Using this factorization results in a search space that is substantially smaller than its convex counterpart, and moreover, in a single use of top- r SVD during the entire execution algorithm. Bypassing these drawbacks, together with accelerating motions, allows us to estimate quantum states of larger qubit systems than state-of-the-art algorithms.

Full QST using non-convex optimization. The use of non-convex algorithms in QST was studied in the context of full tomography as well. By “full tomography” we refer to the situation where an informationally complete measurement is performed, so that the input data to the algorithm is of size 4^n . The exponential scaling of the data size restrict the applicability of full tomography to relatively small system sizes. In this setting non-convex algorithms which work in the factored space were studied [8, 71, 91, 34, 84, 84]. Except of the work [34], we are not aware of theoretical results on the convergence of the proposed algorithm due to the presence of spurious local minima. The authors of [34] characterize the local vs. the global behavior of the objective function under the factorization $\rho = UU^\dagger$ and discuss how existing methods fail due to improper stopping criteria or due to the lack of algorithmic convergence results. Their work highlights the lack of rigorous convergence results of non-convex algorithms used in full quantum state tomography. There is no available implementation publicly available for these methods as well.

Full QST using convex optimization. Despite the non-scalability of full QST, and the limitation of convex optimization, a lot of research was devoted to this topic. Here, we mention only a few notable results that extend the applicability of full QST using specific techniques in convex optimization. Ref [81] shows that for given measurement schemes the solution for the maximum likelihood is given by a linear inversion scheme, followed by a projection onto the set of density matrices. More recently, the authors of [39] used a combination of the techniques of [81] with the sparsity of the Pauli matrices and the use of GPUs to perform a full QST of 14 qubits. While pushing the limit of full QST using convex optimization, obtaining full tomographic *experimental* data for more than a dozen qubits is significantly time-intensive. Moreover, this approach is highly centralized, in comparison to our approach that can be distributed. Using the sparsity pattern property of the Pauli matrices and GPUs is an excellent candidate approach to further enhance the performance of non-convex compressed sensing QST.

QST using neural networks. Deep neural networks are ubiquitous, with many applications to science and industry. Recently, [86, 10, 87, 29] show how machine learning and neural networks can be used to perform QST, driven by experimental data. The neural network architecture used is based on restricted Boltzmann machines (RBMs) [83], which feature a visible and a hidden layer of stochastic binary neurons, fully connected with weighted edges. Test cases considered include reconstruction of W state, magnetic observables

of local Hamiltonians, the unitary dynamics induced by Hamiltonian evolution. Comparison results are provided in the Main Results section.

QST for Matrix Product States (MPS). This is the case of highly structured quantum states where the state is well-approximated by a MPS of low bond dimension [21, 55]. The idea behind this approach is, in order to overcome exponential bottlenecks in the general QST case, we require highly structured subsets of ground truth states, similar to the assumptions made in compressed sensing QST. MPS QST is considered an alternative approach to reduce the computational and storage complexity of QST.

Direct fidelity estimation. Rather than focusing on entrywise estimation of density matrices, the direct fidelity estimation procedure focuses on checking how close is the state of the system to a target state, where closeness is quantified by the fidelity metric. Classic techniques require up to $2^n/\epsilon^4$ number of samples, where ϵ denotes the accuracy of the fidelity term, when considering a general quantum state [42], but can be reduced to almost dimensionality-free $1/\epsilon^2$ number of samples for specific cases, such as stabilizer states [27, 22, 47]. Shadow tomography is considered as an alternative and generalization of this technique [1, 2]; however, as noted in [42], the procedure in [1, 2] requires exponentially long quantum circuits that act collectively on all the copies of the unknown state stored in a quantum memory, and thus has not been implemented fully on real quantum machines. A recent neural network-based implementation of such indirect QST learning methods is provided here [80].

The work in [70, 42], goes beyond simple fidelity estimation, and utilizes random single qubit rotations to learn a minimal sketch of the unknown quantum state by which one that can predict arbitrary linear function of the state. Such methods constitute a favorable alternative to QST as they do not require number of samples that scale polynomially with the dimension; however, this, in turn, implies that these methods cannot be used in general to estimate the density matrix itself.

Methods

MiFGD algorithm. Algorithm 1 contains the details of the Momentum-Inspired Factored Gradient Descent.

Algorithm 1 Momentum-Inspired Factored Gradient Descent

Input: \mathcal{A} , y , r , μ , and # iterations J .
Set U_0 randomly or as in (11).
Set $Z_0 = U_0$.
Set η as in (12).
for $i = 0$ to $J - 1$ **do**
 $U_{i+1} = Z_i - \eta \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z_i$
 $Z_{i+1} = U_{i+1} + \mu(U_{i+1} - U_i)$
end for
Output: $\rho = U_J U_J^\dagger$

The initial point U_0 is either randomly selected [13, 74], or set according to Lemma 4 in [54] to:

$$\rho_0 = U_0 U_0^\dagger = \Pi_C \left(\frac{-1}{1+\delta_{2r}} \cdot \nabla f(0) \right) = \frac{1}{1+\delta_{2r}} \Pi_C \left(\sum_{i=1}^m y_i A_i \right) \quad (11)$$

where $\Pi_C(\cdot)$ is the projection onto the set of PSD matrices, $\delta_{2r} \in (0, 1)$ is the RIP constant and $\nabla f(0)$ denoted the gradient of f evaluated at the all-zero matrix. Since computing the RIP constants is NP-hard, in practice we compute U_0 through $\rho_0 = \frac{-1}{\hat{L}} \Pi_C \left(\sum_{i=1}^m y_i A_i \right)$, where $\hat{L} \in (1, 11/10]$, see Theorem 2 below. Compared to randomly selecting U_0 , Eq. 11 involves a gradient descent computation and a top- r eigenvalue calculation. As for the step size in algorithm 1, following Lemma 3 below, it is set to

$$\eta = \frac{1}{4((1+\delta_{2r})\|Z_0 Z_0^\dagger\|_2 + \|\mathcal{A}^\dagger(\mathcal{A}(Z_0 Z_0^\dagger) - y)\|_2)}, \quad (12)$$

where $Z_0 = U_0$. Here as well, in practice we replace δ_{2r} by \hat{L} . The step size η remains constant at every iteration step of the algorithm, and requires only two top-eigenvalue computations to calculate the spectral

norms $\|Z_0 Z_0^\dagger\|_2$ and $\|\mathcal{A}^\dagger(\mathcal{A}(Z_0 Z_0^\dagger - y))\|_2$. These computations can be efficiently implemented by any off-the-shelf eigenvalue solver, such as the Power Method or the Lanczos method.

Theorem 2 (Iteration invariant and convergence rate). *Assume that \mathcal{A} satisfies the RIP with constant $\delta_{2r} \leq 1/10$. Let U_0 be such that $\min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \leq \frac{\sqrt{\sigma_r(\rho^*)}}{10^3 \sqrt{\kappa \tau(\rho^*)}}$, where $\kappa := \frac{1+\delta_{2r}}{1-\delta_{2r}}$, $\tau(\rho) := \frac{\sigma_1(\rho)}{\sigma_r(\rho)}$ for rank- r ρ , and $\sigma_i(\rho)$ is the i th singular value of ρ . Set η according to (12) and the momentum parameter $\mu = \frac{\varepsilon}{2 \cdot 10^3 r \tau(\rho^*) \sqrt{\kappa}}$ for user-defined $\varepsilon \in (0, 1]$. For $y = \mathcal{A}(\rho^*)$, where $\text{rank}(\rho^*) = r$, if $\tau(\rho^*) \leq 50$, then MiFGD returns a solution such that*

$$\begin{aligned} \min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F &\leq \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \left((1 + \xi(1 + 2\mu)) + \frac{\xi(1 + \xi(1 + \frac{1}{10^3}))}{1 - \xi(1 + \frac{1}{10^3})} \cdot \frac{\varepsilon}{4} \right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}} \\ &\quad + \frac{\xi \cdot \varepsilon}{4(1 - \xi(1 + \frac{1}{10^3}))} \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}} \\ &\lesssim \alpha c^{J+1} \cdot \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F + O(\mu). \end{aligned} \quad (13)$$

where $c := |\lambda_1| < 1$, $\alpha := \frac{2}{|\lambda_1| - |\lambda_2|} \cdot \left((1 + \xi(1 + 2\mu)) + \frac{\xi(1 + \xi(1 + \frac{1}{10^3}))}{1 - \xi(1 + \frac{1}{10^3})} \cdot \frac{\varepsilon}{4} \right)$, $\lambda_1 > \lambda_2$ are the eigenvalues of $A = \begin{bmatrix} \xi|1 + \mu| & \xi|\mu| \\ 1 & 0 \end{bmatrix}$, and $\xi := \sqrt{1 - \frac{0.393}{\kappa \tau(\rho^*)}}$. That is, the algorithm has a linear convergence rate in iterate distances (first term on the right hand side of Eq. (13)), up to a constant proportional to the momentum parameter μ (second term on the right hand side of Eq. (13)).

Further, U_i satisfies $\min_{R \in \mathcal{O}} \|U_i - U^* R\|_F \leq \frac{\sqrt{\sigma_r(\rho^*)}}{10^3 \sqrt{\kappa \tau(\rho^*)}}$, for each $i \geq 1$, i.e., the algorithm preserves the per-iteration conditions to continue decreasing the distance to the optimum.

The interpretation of the theorem is that the right hand side of Eq. (13) depends on the initial distance $\min_{R \in \mathcal{O}} \|U_0 - U^* R\|_2$, as in convex optimization, where the two parameters c, α appear as contraction constants. Crucially, since $c < 1$ the product αc^J goes exponentially fast to zero. Thus, assuming that the initial point U_0 is close enough to the optimum, as stated in the theorem, MiFGD decreases its distance to U^* with a linear rate, up to an “error” level that depends on the momentum parameter μ and it is bounded by $\frac{1}{2 \cdot 10^3 r \tau(\rho^*) \sqrt{\kappa}}$.

Theorem 2 requires a strong assumption on the momentum parameter μ , which depends on quantities that might not be known *a priori* for general problems. However, we note that for the special case of QST, we know these quantities exactly: r is the rank of density matrix—thus, for pure states this value is equal to $r = 1$; $\tau(\rho^*)$ is the (rank-restricted) condition number of the density matrix ρ —for pure states, $\tau(\rho^*) = \frac{\sigma_1(\rho)}{\sigma_r(\rho)} = \frac{\sigma_1(\rho)}{\sigma_1(\rho)} = 1$; and, finally, κ is the condition number of the sensing map, where, given the constraint $\delta_{2r} \leq 1/10$, leads to the following bound: $\kappa \leq 11/9$. The above lead to a momentum value $\mu \approx \varepsilon/2211$. However, as we show in the numerical experiments, the theory is conservative; much larger values of μ lead to stable, improved performance. Finally, the bound on the condition number in Theorem 2 is not strict, and comes out of the analysis we follow; we point the reader to similar assumptions made where $\tau(\rho^*)$ is assumed constant $O(1)$ [59].

The detailed proof is provided in the supplementary material. To the best of our knowledge, this is the first proof for momentum-inspired factorization technique, under common assumptions: both regarding the problem setting, and the assumptions made for its completion. The proof differs from state of the art proofs for non-accelerated factored gradient descent: due to the inclusion of the memory term, three different terms $-U_{i+1}, U_i, U_{i-1}$ —need to be handled simultaneously. Further, the proof differs from recent proofs on non-convex, but non-factored, gradient descent methods, as in [49]: the distance metric over rotations $\min_{R \in \mathcal{O}} \|Z_i - U^* R\|_F$, where Z_i includes estimates from two steps in history, is not amenable to simple triangle inequality bounds, and a careful analysis is required. The analysis requires the design of two-dimensional dynamical systems, where we require to characterize and bound the eigenvalues of a 2×2 contraction matrix.

Acknowledgements

Anastasios Kyrillidis and Amir Kalev acknowledge funding by the NSF (CCF-1907936). Anastasios Kyrillidis thanks Mads Mikkelsen for his performance in “Jagten” and “Druk”.

Data availability

The empirical results were obtained via synthetic and real experiments; the algorithm’s implementation is available at <https://github.com/gidiko/MiFGD>.

Competing interests

The authors declare no competing financial or non-financial interests.

Author contribution

All authors have made substantial contributions to the paper: design of the work, drafting the manuscript, final approval and accountability for all aspects of the work.

References

- [1] Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 325–338, 2018.
- [2] Scott Aaronson and Guy N Rothblum. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 322–333, 2019.
- [3] P. Achuthan and K. Venkatesan. General principles of quantum mechanics. *Handbuch der Physik*, 5(Part 1), 1958.
- [4] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.
- [5] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.
- [6] Robert Alicki, Sl/awomir Rudnicki, and Sl/awomir Sadowski. Symmetry properties of product states for the system of n n -level atoms. *Journal of mathematical physics*, 29(5):1158–1162, 1988.
- [7] Joseph B Altepeter, Daniel FV James, and Paul G Kwiat. 4 qubit quantum state tomography. In *Quantum state estimation*, pages 113–145. Springer, 2004.
- [8] K. Banaszek, G. M. D’Ariano, M. G. A. Paris, and M. F. Sacchi. Maximum-likelihood estimation of the density matrix. *Physical Review A*, 61(1):010304, 1999.
- [9] Konrad Banaszek, Marcus Cramer, and David Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.
- [10] Matthew JS Beach, Isaac De Vlught, Anna Golubeva, Patrick Huembeli, Bohdan Kulchytskyy, Xiuzhe Luo, Roger G Melko, Ejaaz Merali, and Giacomo Torlai. Qucumber: wavefunction reconstruction with neural networks. *SciPost Physics*, 7(1):009, 2019.
- [11] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

- [12] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016.
- [13] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.
- [14] E. Bolduc, G. Knee, E. Gauger, and J. Leach. Projected gradient descent algorithms for quantum state tomography. *npj Quantum Information*, 3(1):44, 2017.
- [15] S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to Nesterov’s accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.
- [16] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
- [17] E. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.
- [18] Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.
- [19] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.
- [20] Matthias Christandl and Graeme Mitchison. The spectra of quantum states and the kronecker coefficients of the symmetric group. *Communications in mathematical physics*, 261(3):789–797, 2006.
- [21] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu. Efficient quantum state tomography. *Nat. Comm.*, 1:149, 2010.
- [22] Marcus P da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. *Physical Review Letters*, 107(21):210404, 2011.
- [23] Lisandro D Dalcin, Rodrigo R Paz, Pablo A Kler, and Alejandro Cosimo. Parallel distributed computing using python. *Advances in Water Resources*, 34(9):1124–1139, 2011.
- [24] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- [25] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *arXiv preprint arXiv:1910.06343*, 2019.
- [26] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.
- [27] Steven T Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Physical review letters*, 106(23):230501, 2011.
- [28] The MPI Forum. Mpi: A message passing interface. In *Proceedings of the 1993 ACM/IEEE Conference on Supercomputing*, Supercomputing ’93, page 878–883, New York, NY, USA, 1993. Association for Computing Machinery.
- [29] Xun Gao and Lu-Ming Duan. Efficient representation of quantum many-body states with deep neural networks. *Nature communications*, 8(1):1–6, 2017.
- [30] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

- [31] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [32] G. Goh. Why momentum really works. *Distill*, 2017.
- [33] D. Gonçalves, M. Gomes-Ruggiero, and C. Lavor. A projected gradient method for optimization over density matrices. *Optimization Methods and Software*, 31(2):328–341, 2016.
- [34] D. Gonçalves, M. Gomes-Ruggiero, C. Lavor, O. J. Farias, and P. Ribeiro. Local solutions of maximum likelihood estimation in quantum state tomography. *Quantum Information & Computation*, 12(9-10):775–790, 2012.
- [35] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
- [36] D. Gross, Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.
- [37] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, 2002.
- [38] R. Horn and Ch. Johnson. *Matrix analysis*. Cambridge university press, 1990.
- [39] Zhibo Hou, Han-Sen Zhong, Ye Tian, Daoyi Dong, Bo Qi, Li Li, Yuanlong Wang, Franco Nori, Guo-Yong Xiang, Chuan-Feng Li, et al. Full reconstruction of a 14-qubit state within four hours. *New Journal of Physics*, 18(8):083036, 2016.
- [40] Ya-Ping Hsieh, Yu-Chun Kao, Rabeeh Karimi Mahabadi, Yurtsever Alp, Anastasios Kyrillidis, and Volkan Cevher. A non-euclidean gradient descent framework for non-convex matrix factorization. Technical report, Institute of Electrical and Electronics Engineers, 2017.
- [41] Zhilin Hu, Kezhi Li, Shuang Cong, and Yaru Tang. Reconstructing pure 14-qubit quantum states in three hours using compressive sensing. *IFAC-PapersOnLine*, 52(11):188 – 193, 2019. 5th IFAC Conference on Intelligent Control and Automation Sciences ICONS 2019.
- [42] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *arXiv preprint arXiv:2002.08953*, 2020.
- [43] Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.
- [44] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.
- [45] Miroslav Ježek, Jaromír Fiurášek, and Zdeněk Hradil. Quantum inference of states and processes. *Physical Review A*, 68(1):012305, 2003.
- [46] A. Kalev, R. Kosut, and I. Deutsch. Quantum tomography protocols with positivity are compressed sensing protocols. *NPJ Quantum Information*, 1:15018, 2015.
- [47] Amir Kalev, Anastasios Kyrillidis, and Norbert M Linke. Validating and certifying stabilizer states. *Physical Review A*, 99(4):042337, 2019.
- [48] Michael Keyl and Reinhard F Werner. Estimating the spectrum of a density operator. In *Asymptotic Theory Of Quantum Statistical Inference: Selected Papers*, pages 458–467. World Scientific, 2005.
- [49] R. Khanna and A. Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. *arXiv preprint arXiv:1712.09379*, 2017.
- [50] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [51] Martin Kliesch, Richard Kueng, Jens Eisert, and David Gross. Guaranteed recovery of quantum processes from few measurements. *Quantum*, 3:171, 2019.
- [52] A. Kyrillidis and V. Cevher. Recipes on hard thresholding methods. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*, pages 353–356. IEEE, 2011.
- [53] A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. *Journal of mathematical imaging and vision*, 48(2):235–265, 2014.
- [54] A. Kyrillidis, A. Kalev, D. Park, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable quantum state tomography via non-convex methods. *npj Quantum Information*, 4(36), 2018.
- [55] BP Lanyon, C Maier, Milan Holzäpfel, Tillmann Baumgratz, C Hempel, P Jurcevic, Ish Dhand, AS Buyskikh, AJ Daley, Marcus Cramer, et al. Efficient tomography of a quantum many-body system. *Nature Physics*, 13(12):1158–1162, 2017.
- [56] J. Lee, M. Simchowitz, M. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.
- [57] Kiryung Lee and Yoram Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. *arXiv preprint arXiv:0903.4742*, 2009.
- [58] Kiryung Lee and Yoram Bresler. Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.
- [59] Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *International Conference on Artificial Intelligence and Statistics*, 2019.
- [60] Y.-K. Liu. Universal low-rank matrix recovery from Pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646, 2011.
- [61] Zhang Liu and Lieven Vandenbergh. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.
- [62] L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für mathematik*, 79(4):303–306, 1975.
- [63] Masoud Mohseni, AT Rezakhani, and DA Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.
- [64] S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.
- [65] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.
- [66] Ryan O’Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 899–912, 2016.
- [67] B. O’Donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
- [68] B. O’Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016.
- [69] B. O’Donoghue, E. Chu, N. Parikh, and S. Boyd. SCS: Splitting conic solver, version 2.1.2. <https://github.com/cvxgrp/scs>, November 2019.

- [70] Marco Pains and Amir Kalev. An approximate description of quantum states. *arXiv preprint arXiv:1910.10543*, 2019.
- [71] M. Paris, G. D’Ariano, and M. Sacchi. Maximum-likelihood method in quantum estimation. In *AIP Conference Proceedings*, volume 568, pages 456–467. AIP, 2001.
- [72] Dohyung Park, Anastasios Kyrillidis, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.
- [73] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions to matrix problems, efficiently and provably. *arXiv preprint arXiv:1606.03168*, 2016.
- [74] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.
- [75] QISKit Development Team. Qiskit.
- [76] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [77] C. Riofrío, D. Gross, S.T. Flammia, T. Monz, D. Nigg, R. Blatt, and J. Eisert. Experimental quantum compressed sensing for a seven-qubit system. *Nature Communications*, 8, 2017.
- [78] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045, 2008.
- [79] Jiangwei Shang, Zhengyun Zhang, and Hui Khoo Ng. Superfast maximum-likelihood reconstruction for quantum tomography. *Phys. Rev. A*, 95:062336, Jun 2017.
- [80] A. Smith, Johnnie Gray, and M. Kim. Efficient approximate quantum state tomography with basis dependent neural-networks. *arXiv preprint arXiv:2009.07601*, 2020.
- [81] J. A. Smolin, J. M. Gambetta, and G. Smith. Efficient method for computing the maximum-likelihood quantum state from measurements with additive Gaussian noise. *Physical review letters*, 108(7):070502, 2012.
- [82] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.
- [83] Ilya Sutskever, Geoffrey E Hinton, and Graham W Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in neural information processing systems*, pages 1601–1608, 2009.
- [84] Y. S. Teo, J. Řeháček, and Z. Hradil. Informationally incomplete quantum tomography. *Quantum Measurements and Quantum Metrology*, 1, 2013.
- [85] T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- [86] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nat. Phys.*, 14:447–450, May 2018.
- [87] Giacomo Torlai and Roger Melko. Machine-learning quantum states in the NISQ era. *Annual Review of Condensed Matter Physics*, 11, 2019.
- [88] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 964–973. JMLR. org, 2016.

- [89] L. Vandenberghe. The CVXOPT linear and quadratic cone program solvers. Online: <http://cvxopt.org/documentation/coneprog.pdf>, 2010.
- [90] K Vogel and H Risken. Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. *Physical Review A*, 40(5):2847, 1989.
- [91] J. Řeháček, Z. Hradil, E. Knill, and A. I. Lvovsky. Diluted maximum-likelihood algorithm for quantum tomography. *Phys. Rev. A*, 75:042108, 2007.
- [92] Andrew E Waters, Aswin C Sankaranarayanan, and Richard Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.
- [93] K. Williams. The n -th power of a 2×2 matrix. *Mathematics Magazine*, 65(5):336–336, 1992.
- [94] Peng Xu, Bryan He, Christopher De Sa, Ioannis Mitliagkas, and Chris Re. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 58–67, 2018.
- [95] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.
- [96] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

Appendix

IBM Quantum system experiments: GHZ₋(6) circuit, 2048 shots

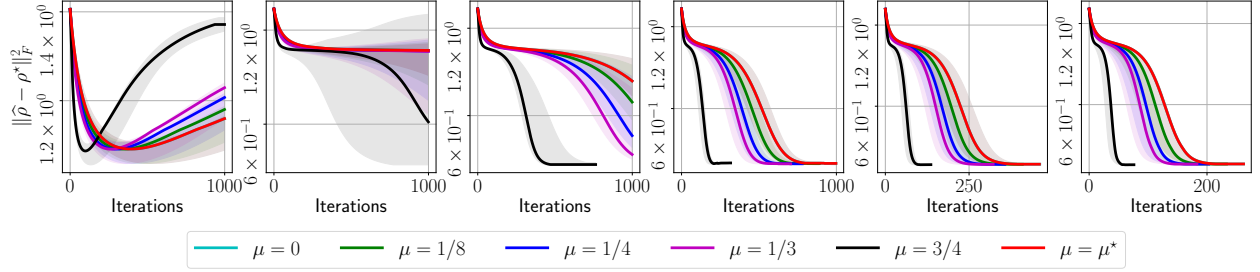


Figure 8: Target error list plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments.

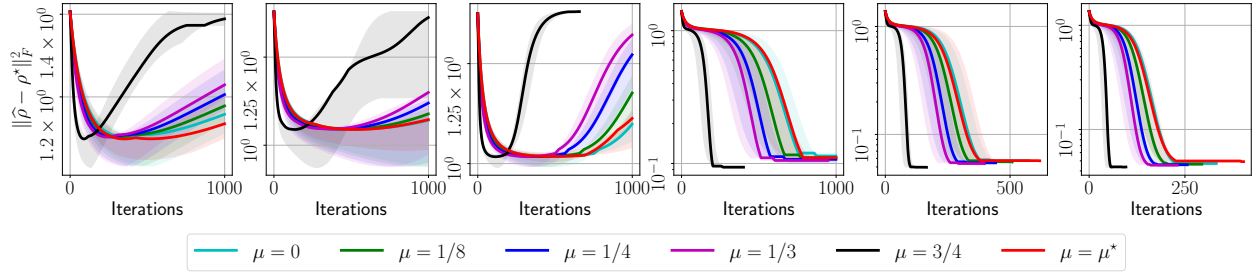


Figure 9: Target error list plots for reconstructing GHZ₋(6) circuit using synthetic measurements from IBM's quantum simulator.

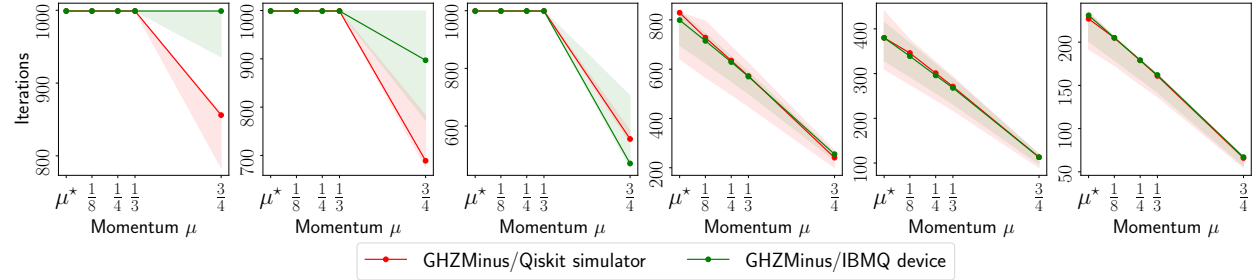


Figure 10: Convergence iteration plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

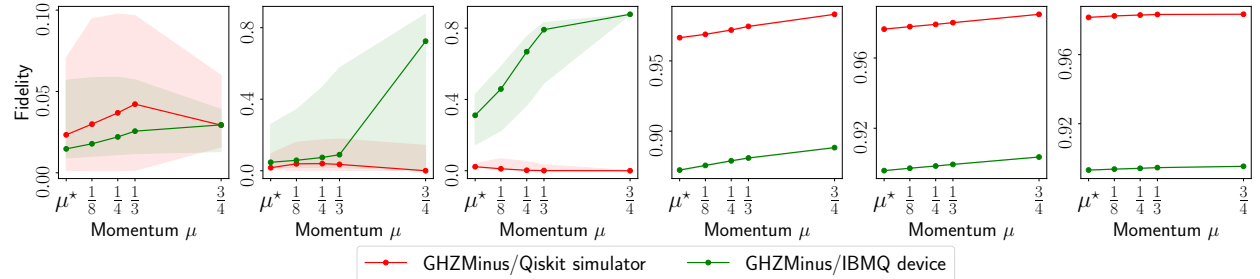


Figure 11: Fidelity list plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

IBM Quantum system experiments: GHZ₋(6) circuit, 8192 shots

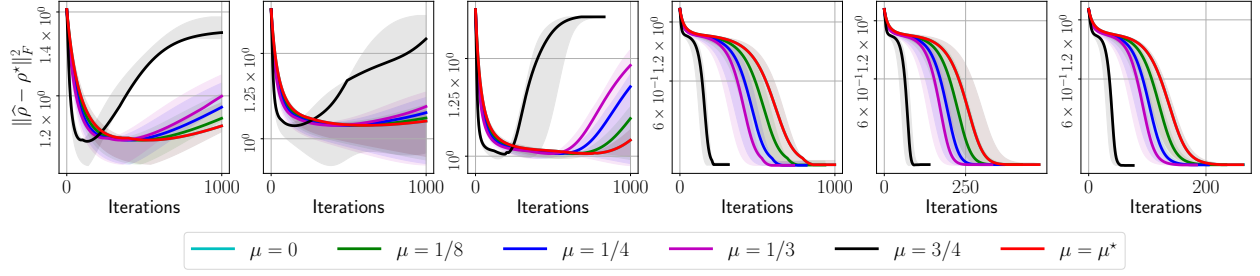


Figure 12: Target error list plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments.

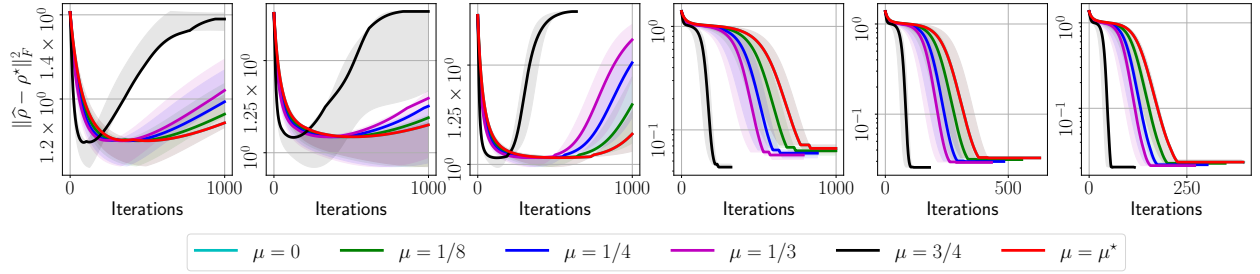


Figure 13: Target error list plots for reconstructing GHZ₋(6) circuit using synthetic measurements from IBM's quantum simulator.

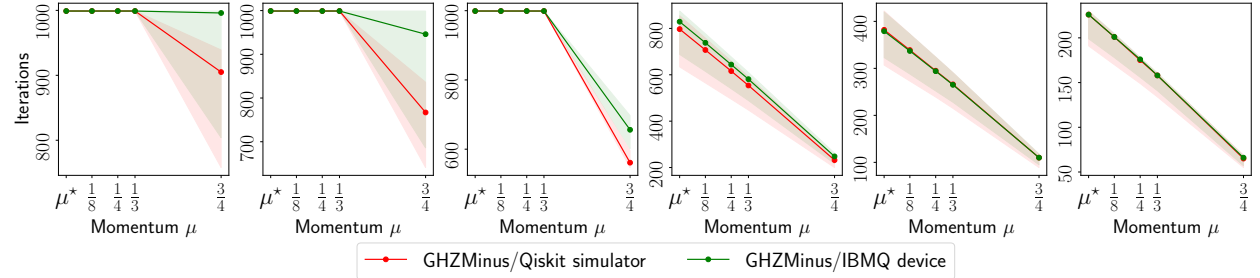


Figure 14: Convergence iteration plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

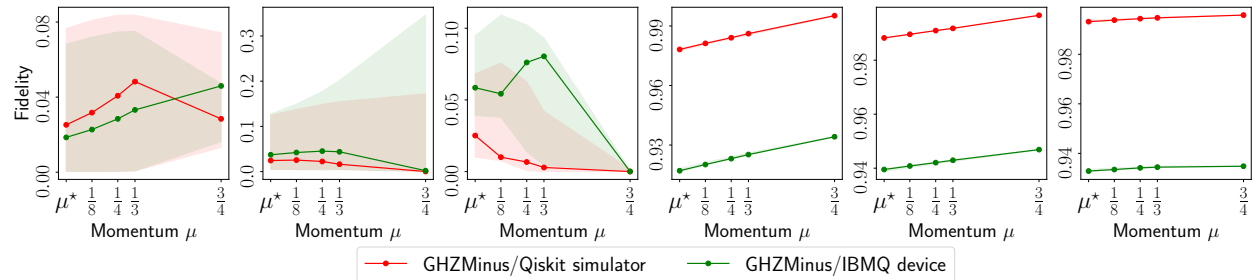


Figure 15: Fidelity list plots for reconstructing GHZ₋(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

IBM Quantum system experiments: GHZ₋(8) circuit, 2048 shots

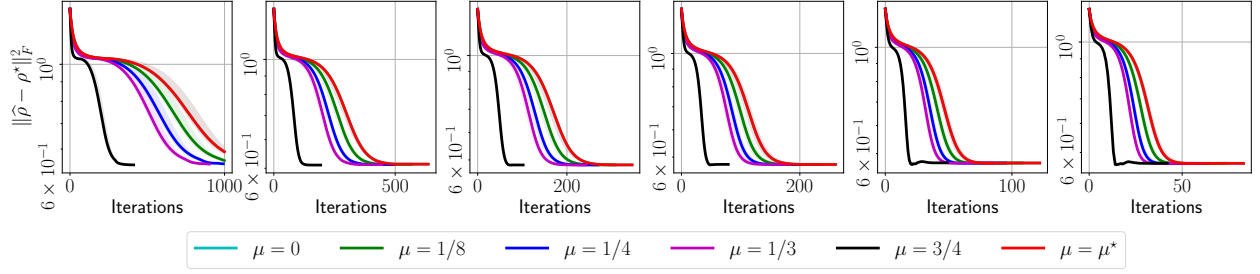


Figure 16: Target error list plots for reconstructing GHZ₋(8) circuit using real measurements from IBM Quantum system experiments.

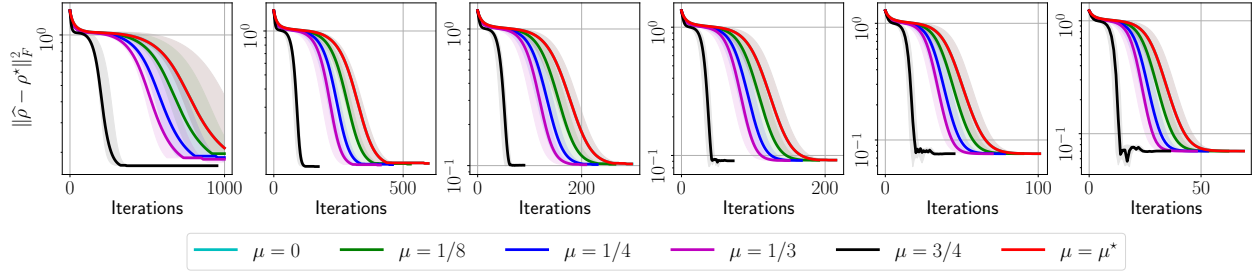


Figure 17: Target error list plots for reconstructing GHZ₋(8) circuit using synthetic measurements from IBM's quantum simulator.

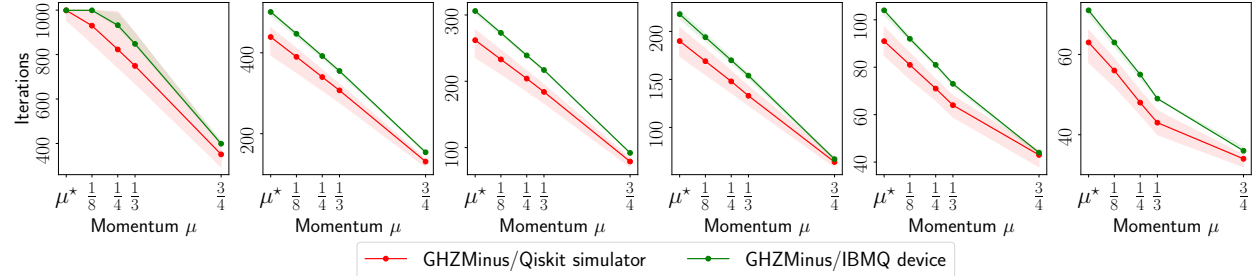


Figure 18: Convergence iteration plots for reconstructing GHZ₋(8) circuit using using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

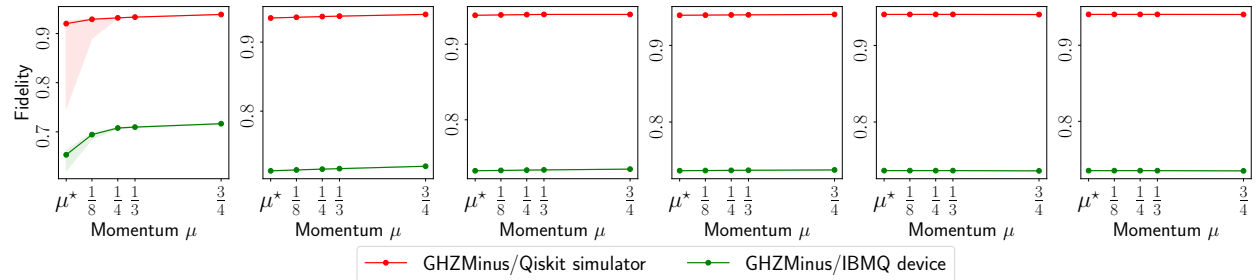


Figure 19: Fidelity list plots for reconstructing GHZ₋(8) circuit using using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

IBM Quantum system experiments: GHZ₋(8) circuit, 4096 shots

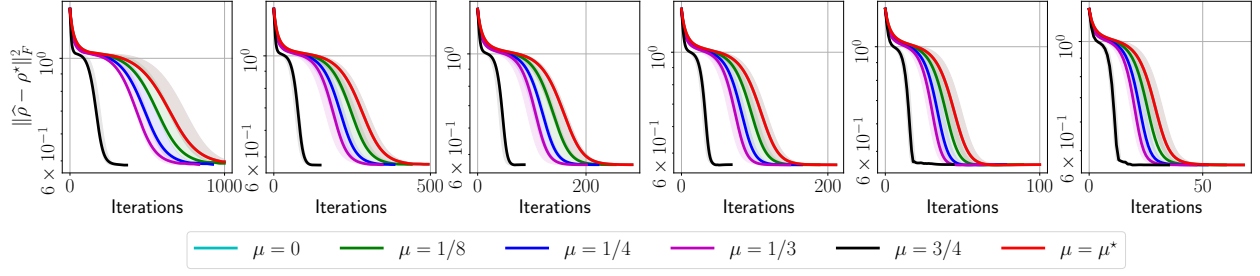


Figure 20: Target error list plots for reconstructing GHZ₋(8) circuit using real measurements from IBM Quantum system experiments.

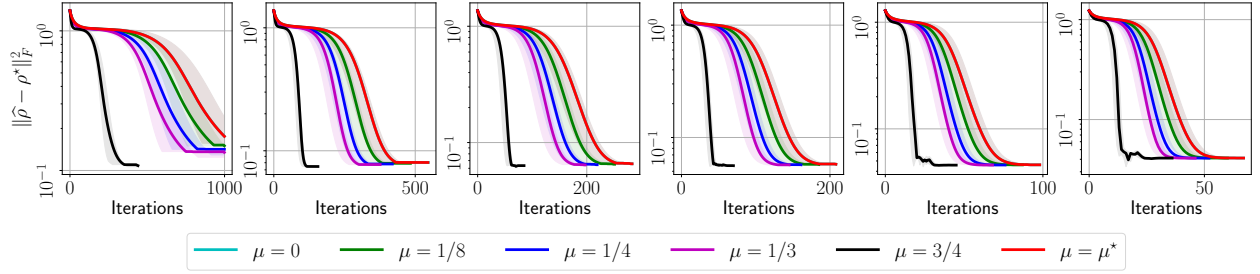


Figure 21: Target error list plots for reconstructing GHZ₋(8) circuit using synthetic measurements from IBM's quantum simulator.

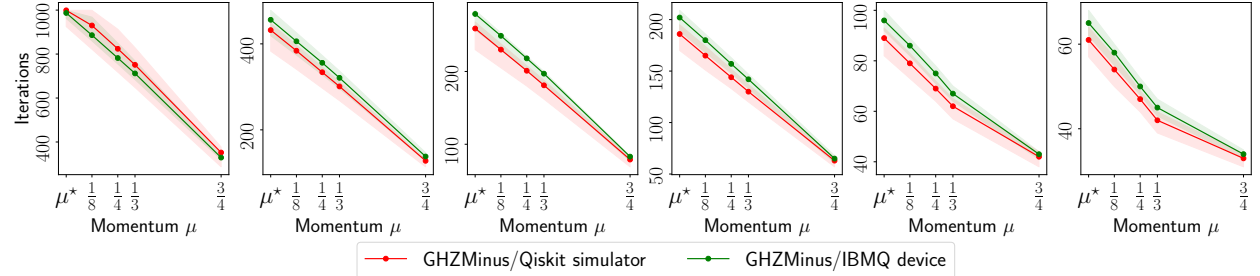


Figure 22: Convergence iteration plots for reconstructing GHZ₋(8) circuit using using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

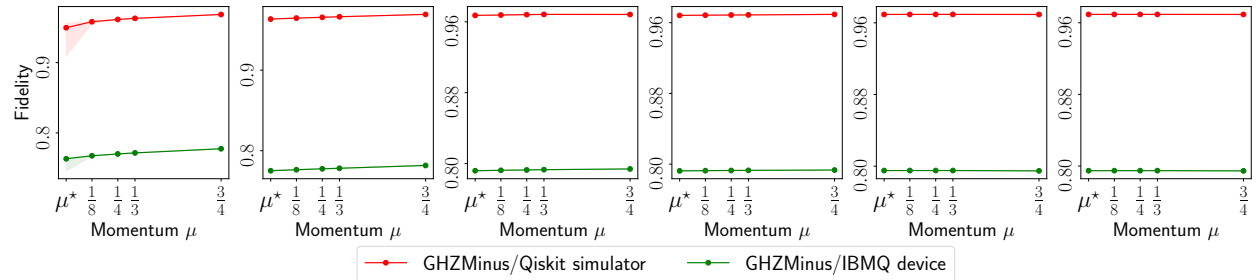


Figure 23: Fidelity list plots for reconstructing GHZ₋(8) circuit using using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

IBM Quantum system experiments: Hadamard(6) circuit, 8192 shots

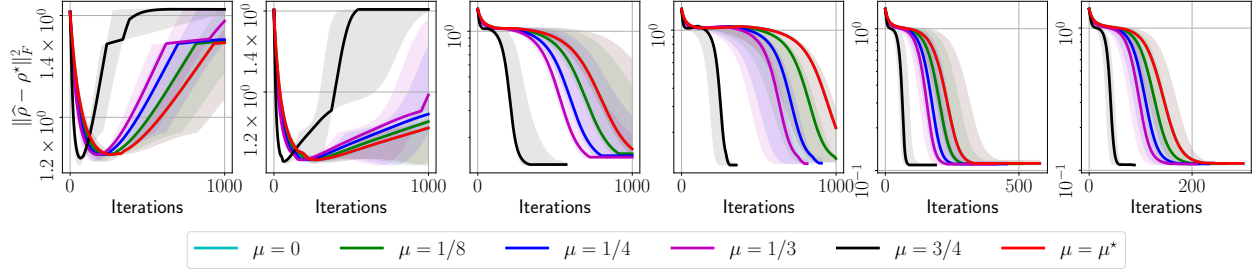


Figure 24: Target error list plots for reconstructing Hadamard(6) circuit using real measurements from IBM Quantum system experiments.

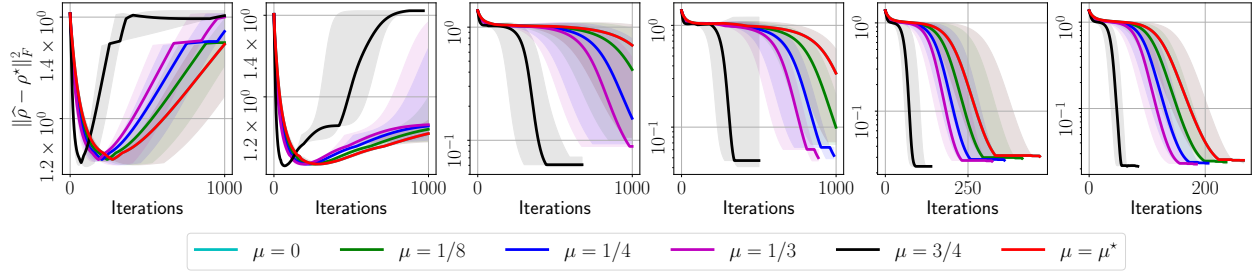


Figure 25: Target error list plots for reconstructing Hadamard(6) circuit using synthetic measurements from IBM's quantum simulator.

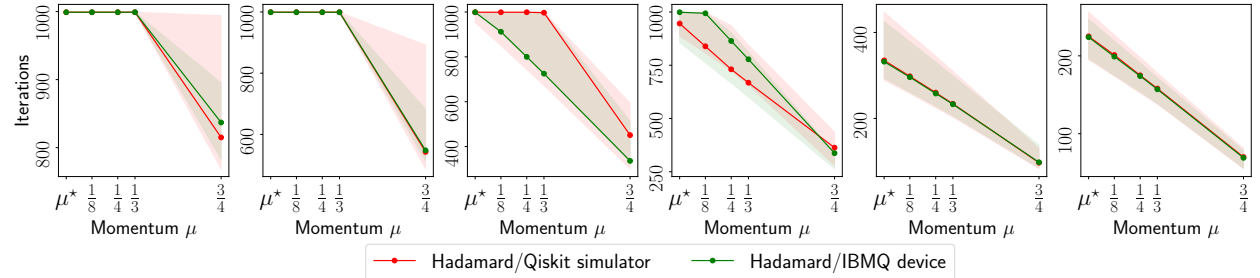


Figure 26: Convergence iteration plots for reconstructing Hadamard(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation.

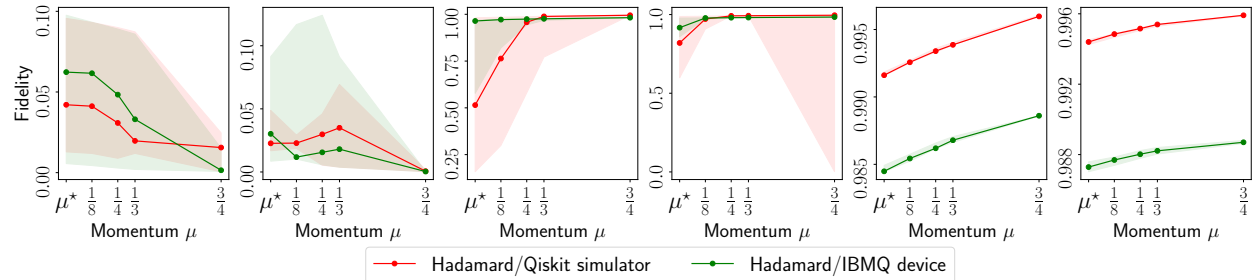


Figure 27: Fidelity list plots for reconstructing Hadamard(6) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

IBM Quantum system experiments: Hadamard(8) circuit, 4096 shots

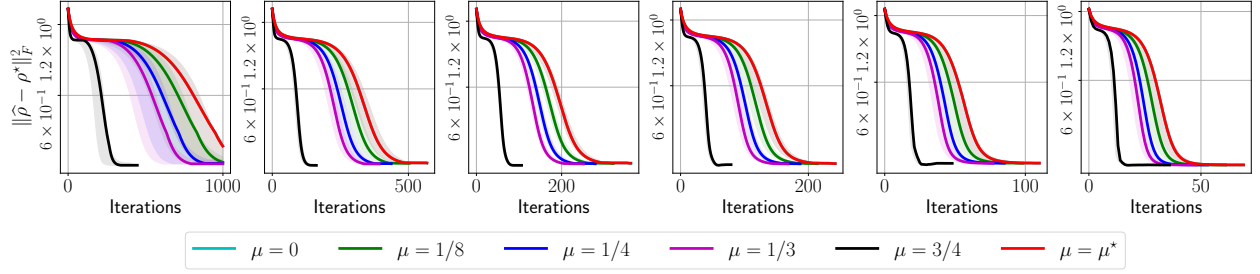


Figure 28: Target error list plots for reconstructing Hadamard(8) circuit using real measurements from IBM Quantum system experiments.

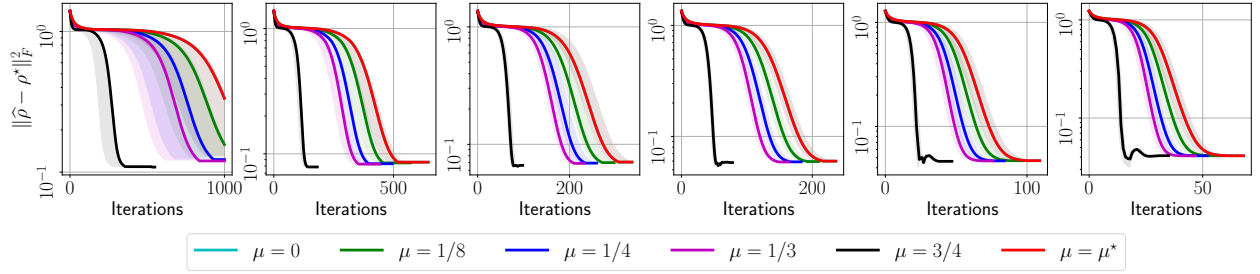


Figure 29: Target error list plots for reconstructing Hadamard(8) circuit using synthetic measurements from IBM's quantum simulator.

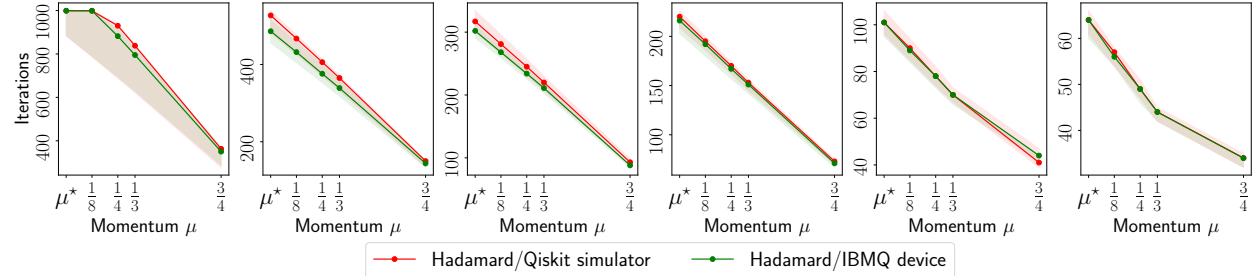


Figure 30: Convergence iteration plots for reconstructing Hadamard(8) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation.

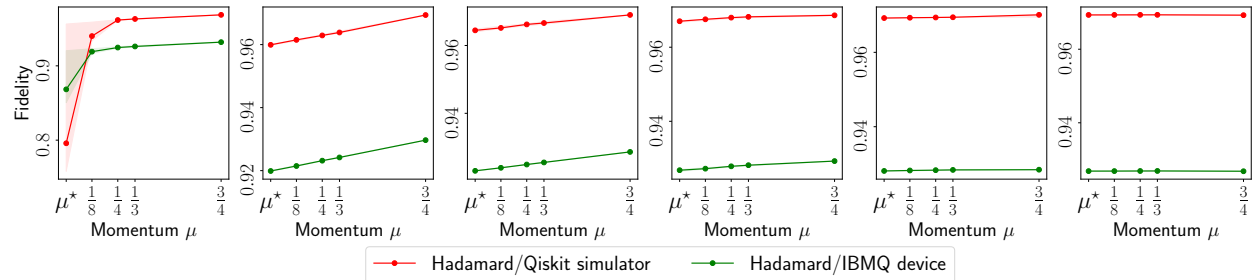


Figure 31: Fidelity list plots for reconstructing Hadamard(8) circuit using real measurements from IBM Quantum system experiments and synthetic measurements from Qiskit simulation experiments.

Synthetic experiments for $n = 12$

We compare MiFGD with *i*) the Matrix ALPS framework [53], a state of the art projected gradient descent algorithm, and an optimized version of matrix iterative hard thresholding, operating on the full matrix variable ρ , with adaptive step size η (we note that this algorithm has outperformed most of the schemes that work on the original space ρ ; see [53]); *ii*) the plain Procrustes Flow/FGD algorithm [88, 12, 54], where we use the step size as reported in [12], since the later has reported better performance than vanilla Procrustes Flow. We note that the Procrustes Flow/FGD algorithm is similar to our algorithm without acceleration. Further, the original Procrustes Flow/FGD algorithm relies on performing many iterations in the original space ρ as an initialization scheme, which is often prohibitive as the problem dimensions grow. Both for our algorithm and the plain Procrustes Flow/FGD scheme, we use random initialization.

To properly compare the algorithms in the above list, we pre-select a common set of problem parameters. We fix the dimension $d = 4096$ (equivalent to $n = 12$ qubits), and the rank of the optimal matrix $\rho^* \in \mathbb{R}^{d \times d}$ to be $r = 10$ (equivalent to a mixed quantum state reconstruction). Similar behavior has been observed for other values of r , and are omitted. We fix the number of observables m to be $m = c \cdot d \cdot r$, where $c \in \{3, 5\}$. In all algorithms, we fix the maximum number of iterations to 4000, and we use the same stopping criterion: $\|\rho_{i+1} - \rho_i\|_F / \|\rho_i\|_F \leq \text{tol} = 10^{-3}$. For the implementation of MiFGD, we have used the momentum parameter $\mu = \frac{2}{3}$, as well as the theoretical μ value.

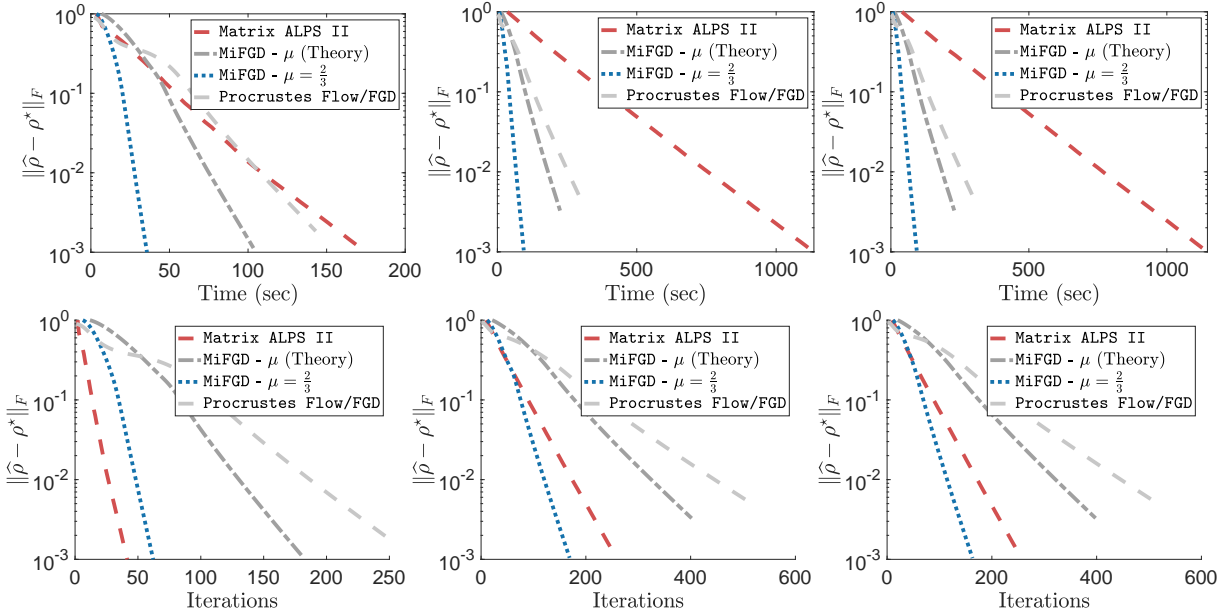


Figure 32: Synthetic example results on low-rank matrix sensing in higher dimensions (equivalent to $n = 12$ qubits). **Top row:** Convergence behavior vs. time elapsed. **Bottom row:** Convergence behavior vs. number of iterations. **Left panel:** $c = 5$, noiseless case; **Center panel:** $c = 3$, noiseless case; **Right panel:** $c = 5$, noisy case, $\|w\|_2 = 0.01$.

The procedure to generate synthetically the data is as follows: The observations y are set to $y = \mathcal{A}(\rho^*) + w$ for some noise vector w ; while the theory holds for the noiseless case, we show empirically that noisy cases are robustly handled by the same algorithm. We use permuted and subsampled noiselets for the linear operator \mathcal{A} [92]. The optimal matrix ρ^* is generated as the multiplication of a tall matrix $U^* \in \mathbb{R}^{d \times r}$ such that $\rho^* = U^* U^{*\top}$, and $\|\rho^*\|_F = 1$, without loss of generality. The entries of U^* are drawn i.i.d. from a Gaussian distribution with zero mean and unit variance. In the noisy case, w has the same dimensions with y , its entries are drawn from a zero mean Gaussian distribution with norm $\|w\|_2 = 0.01$. The random initialization is defined as U_0 drawn i.i.d. from a Gaussian distribution with zero mean and unit variance.

The results are shown in Figure 32. Some notable remarks: *i*) While factorization techniques might take more iterations to converge compared to non-factorized algorithms, the per iteration time complexity is much less, such that overall, factorized gradient descent converges more quickly in terms of total execution

time. *ii)* Our proposed algorithm, *even under the restrictive assumptions on acceleration parameter μ* , performs better than the non-accelerated factored gradient descent algorithms, such as Procrustes Flow. *iii)* Our theory is conservative: using a much larger μ we obtain a faster convergence; the proof for less strict assumptions for μ is an interesting future direction. In all cases, our findings illustrate the effectiveness of the proposed schemes on different problem configurations.

Supporting lemmata

In this section, we present a series of lemmata, used for the main result of the paper.

Lemma 1. Let $U \in \mathbb{C}^{d \times r}$ and $U^* \in \mathbb{C}^{d \times r}$, such that $\|U - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$ for some $R \in \mathcal{O}$, where $\rho^* = U^*U^{*\dagger}$, $\kappa := \frac{1+\delta_{2r}}{1-\delta_{2r}} > 1$, for $\delta_{2r} \leq \frac{1}{10}$, and $\tau(\rho^*) := \frac{\sigma_1(\rho^*)}{\sigma_r(\rho^*)} > 1$. Then:

$$\sigma_1(\rho^*)^{1/2} \left(1 - \frac{1}{10^3}\right) \leq \sigma_1(U) \leq \sigma_1(\rho^*)^{1/2} \left(1 + \frac{1}{10^3}\right) \quad (14)$$

$$\sigma_r(\rho^*)^{1/2} \left(1 - \frac{1}{10^3}\right) \leq \sigma_r(U) \leq \sigma_r(\rho^*)^{1/2} \left(1 + \frac{1}{10^3}\right) \quad (15)$$

Proof. By the fact $\|\cdot\|_2 \leq \|\cdot\|_F$ and using Weyl's inequality for perturbation of singular values [38, Theorem 3.3.16], we have:

$$|\sigma_i(U) - \sigma_i(U^*)| \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3}, \quad 1 \leq i \leq r. \quad (16)$$

Then,

$$-\frac{\sigma_r(\rho^*)^{1/2}}{10^3} \leq \sigma_1(U) - \sigma_1(U^*) \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \Rightarrow \quad (17)$$

$$\sigma_1(\rho^*)^{1/2} - \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \leq \sigma_1(U) \leq \sigma_1(\rho^*)^{1/2} + \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \Rightarrow \quad (18)$$

$$\sigma_1(\rho^*)^{1/2} \left(1 - \frac{1}{10^3}\right) \leq \sigma_1(U) \leq \sigma_1(\rho^*)^{1/2} \left(1 + \frac{1}{10^3}\right). \quad (19)$$

Similarly:

$$-\frac{\sigma_r(\rho^*)^{1/2}}{10^3} \leq \sigma_r(U) - \sigma_r(U^*) \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \Rightarrow \quad (20)$$

$$\sigma_r(\rho^*)^{1/2} - \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \leq \sigma_r(U) \leq \sigma_r(\rho^*)^{1/2} + \frac{\sigma_r(\rho^*)^{1/2}}{10^3} \Rightarrow \quad (21)$$

$$\sigma_r(\rho^*)^{1/2} \left(1 - \frac{1}{10^3}\right) \leq \sigma_r(U) \leq \sigma_r(\rho^*)^{1/2} \left(1 + \frac{1}{10^3}\right). \quad (22)$$

In the above, we used the fact that $\sigma_i(U^*) = \sigma_i(\rho^*)^{1/2}$, for all i , and the fact that $\sigma_i(\rho^*)^{1/2} \geq \sigma_j(\rho^*)^{1/2}$, for $i \leq j$. \square

Lemma 2. Let $U \in \mathbb{C}^{d \times r}$, $U_- \in \mathbb{C}^{d \times r}$, and $U^* \in \mathbb{C}^{d \times r}$, such that $\min_{R \in \mathcal{O}} \|U - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$ and $\min_{R \in \mathcal{O}} \|U_- - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$, where $\rho^* = U^*U^{*\dagger}$, and $\kappa := \frac{1+\delta_{2r}}{1-\delta_{2r}} > 1$, for $\delta_{2r} \leq \frac{1}{10}$, and $\tau(\rho^*) := \frac{\sigma_1(\rho^*)}{\sigma_r(\rho^*)} > 1$. Set the momentum parameter as $\mu = \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \cdot \frac{\varepsilon}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r}$, for $\varepsilon \in (0, 1)$ user-defined. Then,

$$\|Z - U^*R_Z\|_F \leq \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}. \quad (23)$$

Proof. Let $R_U \in \arg \min_{R \in \mathcal{O}} \|U - U^*R\|_F$ and $R_{U_-} \in \arg \min_{R \in \mathcal{O}} \|U_- - U^*R\|_F$. By the definition of the distance function:

$$\|Z - U^*R_Z\|_F = \min_{R \in \mathcal{O}} \|Z - U^*R\|_F = \min_{R \in \mathcal{O}} \|U + \mu(U - U_-) - U^*R\|_F \quad (24)$$

$$= \min_{R \in \mathcal{O}} \|U + \mu(U - U_-) - (1 - \mu + \mu)U^*R\|_F \quad (25)$$

$$\leq |1 + \mu| \cdot \|U - U^*R_U\|_F + |\mu| \cdot \|U_- - U^*R_{U_-}\|_F \quad (26)$$

$$= |1 + \mu| \cdot \|U - U^*R_U\|_F + |\mu| \cdot \|U_- - U^*R_U - U^*R_{U_-} + U^*R_{U_-}\|_F \quad (27)$$

$$= |1 + \mu| \cdot \|U - U^*R_U\|_F + |\mu| \cdot \|(U_- - U^*R_{U_-}) + U^*(R_{U_-} - R_U)\|_F \quad (28)$$

$$\leq |1 + \mu| \cdot \min_{R \in \mathcal{O}} \|U - U^*R\|_F + |\mu| \cdot \min_{R \in \mathcal{O}} \|U_- - U^*R\|_F + |\mu| \cdot \|U^*(R_U - R_{U_-})\|_F \quad (29)$$

$$\leq |1 + \mu| \cdot \min_{R \in \mathcal{O}} \|U - U^*R\|_F + |\mu| \cdot \min_{R \in \mathcal{O}} \|U_- - U^*R\|_F + 2|\mu| \cdot \sigma_1(\rho^*)^{1/2}r \quad (30)$$

$$\stackrel{(i)}{\leq} \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \quad (31)$$

where (i) is due to the fact that $\mu \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \cdot \frac{1}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r}$. We keep μ in the expression, but we use it for clarity for the rest of the proof. \square

Corollary 1. Let $Z \in \mathbb{C}^{d \times r}$ and $U^* \in \mathbb{C}^{d \times r}$, such that $\|Z - U^*R\|_F \leq \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$ for some $R \in \mathcal{O}$, and $\rho^* = U^*U^{*\dagger}$. Then:

$$\sigma_1(\rho^*)^{1/2} \left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right) \leq \sigma_1(Z) \leq \sigma_1(\rho^*)^{1/2} \left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right) \quad (32)$$

$$\sigma_r(\rho^*)^{1/2} \left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right) \leq \sigma_r(Z) \leq \sigma_r(\rho^*)^{1/2} \left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right). \quad (33)$$

Given that $\mu = \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \cdot \frac{\varepsilon}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r} \leq \frac{1}{10^3}$, we get:

$$0.998 \cdot \sigma_1(\rho^*)^{1/2} \leq \sigma_1(Z) \leq 1.0015 \cdot \sigma_1(\rho^*)^{1/2} \quad (34)$$

$$0.998 \cdot \sigma_r(\rho^*)^{1/2} \leq \sigma_r(Z) \leq 1.0015 \cdot \sigma_r(\rho^*)^{1/2}. \quad (35)$$

Proof. The proof follows similar motions as in Lemma 1. \square

Corollary 2. Under the same assumptions of Lemma 1 and Corollary 1, and given the assumptions on μ , we have:

$$\frac{99}{100} \cdot \|\rho^*\|_2 \leq \|ZZ^\dagger\|_2 \leq \frac{101}{100} \cdot \|\rho^*\|_2 \quad (36)$$

$$\frac{99}{100} \cdot \|\rho^*\|_2 \leq \|Z_0 Z_0^\dagger\|_2 \leq \frac{101}{100} \cdot \|\rho^*\|_2 \quad (37)$$

and

$$\frac{99}{101} \cdot \|Z_0 Z_0^\dagger\|_2 \leq \|ZZ^\dagger\|_2 \leq \frac{101}{99} \cdot \|Z_0 Z_0^\dagger\|_2 \quad (38)$$

Proof. The proof is easily derived based on the quantities from Lemma 1 and Corollary 1. \square

Corollary 3. Let $Z \in \mathbb{C}^{d \times r}$ and $U^* \in \mathbb{C}^{d \times r}$, such that $\|Z - U^*R\|_F \leq \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$ for some $R \in \mathcal{O}$, and $\rho^* = U^*U^{*\dagger}$. Define $\tau(W) = \frac{\sigma_1(W)}{\sigma_r(W)}$. Then:

$$\tau(ZZ^\dagger) \leq \beta^2 \tau(\rho^*), \quad (39)$$

where $\beta := \frac{1 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}}{1 - \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}} > 1$, for $\mu \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \cdot \frac{1}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r}$.

Proof. The proof uses the definition of the condition number $\tau(\cdot)$ and the results from Lemma 1 and Corollary 1. \square

Lemma 3. Consider the following three step sizes:

$$\eta = \frac{1}{4 \left((1 + \delta_{2r}) \|Z_0 Z_0^\dagger\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y)\|_2 \right)} \quad (40)$$

$$\hat{\eta} = \frac{1}{4 \left((1 + \delta_{2r}) \|ZZ^\dagger\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Q_Z Q_Z^\dagger\|_2 \right)} \quad (41)$$

$$\eta^* = \frac{1}{4 \left((1 + \delta_{2r}) \|\rho^{*\top}\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y)\|_2 \right)}. \quad (42)$$

Here, $Z_0 \in \mathbb{C}^{d \times r}$ is the initial point, $Z \in \mathbb{C}^{d \times r}$ is the current point, $\rho^* \in \mathbb{C}^{d \times d}$ is the optimal solution, and Q_Z denotes a basis of the column space of Z . Then, under the assumptions that $\min_{R \in \mathcal{O}} \|U - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$, and $\min_{R \in \mathcal{O}} \|Z - U^*R\|_F \leq \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$, and assuming $\mu = \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \cdot \frac{\varepsilon}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r}$, for the user-defined parameter $\varepsilon \in (0, 1)$, we have:

$$\frac{10}{9} \eta \geq \hat{\eta} \geq \frac{10}{10.5} \eta, \quad \text{and} \quad \frac{100}{102} \eta^* \leq \eta \leq \frac{102}{100} \eta^* \quad (43)$$

Proof. The assumptions of the lemma are identical to that of Corollary 2. Thus, we have: $\frac{99}{100} \cdot \|U^*\|_2^2 \leq \|Z\|_2^2 \leq \frac{101}{100} \cdot \|U^*\|_2^2$, $\frac{99}{100} \cdot \|U^*\|_2^2 \leq \|Z_0\|_2^2 \leq \frac{101}{100} \cdot \|U^*\|_2^2$, and $\frac{99}{101} \cdot \|Z_0\|_2^2 \leq \|Z\|_2^2 \leq \frac{101}{99} \cdot \|Z_0\|_2^2$. We focus on the inequality $\hat{\eta} \geq \frac{10}{10.5}\eta$. Observe that:

$$\left\| \mathcal{A}^\dagger \left(\mathcal{A}(ZZ^\dagger) - y \right) Q_Z Q_Z^\dagger \right\|_2 \leq \left\| \mathcal{A}^\dagger \left(\mathcal{A}(ZZ^\dagger) - y \right) \right\|_2 \quad (44)$$

$$= \left\| \mathcal{A}^\dagger \left(\mathcal{A}(ZZ^\dagger) - y \right) - \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) + \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \quad (45)$$

$$\stackrel{(i)}{\leq} (1 + \delta_{2r}) \left\| ZZ^\dagger - Z_0 Z_0^\dagger \right\|_F + \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \quad (46)$$

$$\leq (1 + \delta_{2r}) \left\| ZZ^\dagger - U^* U^{*\dagger} \right\|_F + (1 + \delta_{2r}) \left\| Z_0 Z_0^\dagger - U^* U^{*\dagger} \right\|_F \quad (47)$$

$$+ \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \quad (48)$$

where (i) is due to smoothness via RIP constants of the objective and the fact $\|\cdot\|_2 \leq \|\cdot\|_F$. For the first two terms on the right-hand side, where R_Z is the minimizing rotation matrix for Z , we obtain:

$$\left\| ZZ^\dagger - U^* U^{*\dagger} \right\|_F = \left\| ZZ^\dagger - U^* R_Z Z^\dagger + U^* R_Z Z^\dagger - U^* U^{*\dagger} \right\|_F \quad (49)$$

$$= \left\| (Z - U^* R_Z) Z^\dagger + U^* R_Z (Z - U^* R_Z)^\dagger \right\|_F \quad (50)$$

$$\leq \|Z\|_2 \cdot \|Z - U^* R_Z\|_F + \|U^*\|_2 \cdot \|Z - U^* R_Z\|_F \quad (51)$$

$$\leq (\|Z\|_2 + \|U^*\|_2) \cdot \|Z - U^* R_Z\|_F \quad (52)$$

$$\stackrel{(i)}{\leq} \left(\sqrt{\frac{101}{99}} + \sqrt{\frac{100}{99}} \right) \|Z_0\|_2 \cdot \|Z - U^* R_Z\|_F \quad (53)$$

$$\stackrel{(ii)}{\leq} \left(\sqrt{\frac{101}{99}} + \sqrt{\frac{100}{99}} \right) \|Z_0\|_2 \cdot 0.001 \sigma_r(\rho^*)^{1/2} \quad (54)$$

$$\leq \left(\sqrt{\frac{101}{99}} + \sqrt{\frac{100}{99}} \right) \cdot 0.001 \cdot \sqrt{\frac{100}{99}} \cdot \|Z_0\|_2^2 \quad (55)$$

where (i) is due to the relation of $\|Z\|_2$ and $\|U^*\|_2$ derived above, (ii) is due to Lemma 2. Similarly:

$$\left\| Z_0 Z_0^\dagger - U^* U^{*\dagger} \right\|_F \leq \left(\sqrt{\frac{101}{99}} + \sqrt{\frac{100}{99}} \right) \cdot 0.001 \cdot \sqrt{\frac{100}{99}} \cdot \|Z_0\|_2^2 \quad (56)$$

Using these above, we obtain:

$$\left\| \mathcal{A}^\dagger \left(\mathcal{A}(ZZ^\dagger) - y \right) Q_Z Q_Z^\dagger \right\|_2 \leq \frac{4 \cdot 1(1 + \delta_{2r})}{10^3} \|Z_0 Z_0^\dagger\|_2 + \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \quad (57)$$

Thus:

$$\hat{\eta} = \frac{1}{4 \left((1 + \delta_{2r}) \|Z Z^\dagger\|_2 + \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z Z^\dagger) - y \right) Q_Z Q_Z^\dagger \right\|_2 \right)} \quad (58)$$

$$\geq \frac{1}{4 \left((1 + \delta_{2r}) \frac{101}{99} \|Z_0 Z_0^\dagger\|_2 + \frac{4 \cdot 1(1 + \delta_{2r})}{10^3} \|Z_0 Z_0^\dagger\|_2 + \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \right)} \quad (59)$$

$$\geq \frac{1}{4 \left(\frac{10.5}{10} \cdot (1 + \delta_{2r}) \|Z_0 Z_0^\dagger\|_2 + \left\| \mathcal{A}^\dagger \left(\mathcal{A}(Z_0 Z_0^\dagger) - y \right) \right\|_2 \right)} \quad (60)$$

$$\geq \frac{10}{10.5} \eta \quad (61)$$

Similarly, one gets $\hat{\eta} \leq \frac{10}{9} \eta$.

For the relation between η and η^* , we will prove here the lower bound; similar motions lead to the upper

bound also. By definition, and using the relations in Corollary 2, we get:

$$\eta = \frac{1}{4 \left((1 + \delta_{2r}) \|Z_0 Z_0^\dagger\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y)\|_2 \right)} \quad (62)$$

$$\geq \frac{1}{4 \left((1 + \delta_{2r}) \frac{101}{100} \|\rho^{\star\top}\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y)\|_2 \right)} \quad (63)$$

For the gradient term, we observe:

$$\left\| \mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y) \right\|_2 \leq \left\| \mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y) - \mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y) \right\|_2 + \left\| \mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y) \right\|_2 \quad (64)$$

$$\stackrel{(i)}{=} \left\| \mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y) - \mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y) \right\|_2 \quad (65)$$

$$\stackrel{(ii)}{\leq} (1 + \delta_{2r}) \left\| Z_0 Z_0^\dagger - U^* U^{\star\top} \right\|_F \quad (66)$$

$$\stackrel{(iii)}{\leq} (1 + \delta_{2r}) (\|Z_0\|_2 + \|U^*\|_2) \cdot \|Z - U^* R_Z\|_F \quad (67)$$

$$\stackrel{(iv)}{\leq} (1 + \delta_{2r}) \left(\sqrt{\frac{101}{100}} + 1 \right) \|U^*\|_2 \cdot 0.001 \cdot \|U^*\|_2^2 \quad (68)$$

$$\leq 0.002 \cdot (1 + \delta_{2r}) \|\rho^*\|_2 \quad (69)$$

where (i) is due to $\left\| \mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y) \right\|_2 = 0$, (ii) is due to the restricted smoothness assumption and the RIP, (iii) is due to the bounds above on $\left\| Z_0 Z_0^\dagger - U^* U^{\star\top} \right\|_F$, (iv) is due to the bounds on $\|Z_0\|_2$, w.r.t. $\|U^*\|_2$, as well as the bound on $\|Z - U^* R_Z\|_F$.

Thus, in the inequality above, we get:

$$\eta \geq \frac{1}{4 \left((1 + \delta_{2r}) \frac{101}{100} \|\rho^{\star\top}\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(Z_0 Z_0^\dagger) - y)\|_2 \right)} \quad (70)$$

$$\geq \frac{1}{4 \left((1 + \delta_{2r}) \frac{101}{100} \|\rho^{\star\top}\|_2 + 0.001 \cdot (1 + \delta_{2r}) \|\rho^*\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y)\|_2 \right)} \quad (71)$$

$$\geq \frac{1}{4 \left((1 + \delta_{2r}) \frac{102}{100} \|\rho^{\star\top}\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(\rho^*) - y)\|_2 \right)} \geq \frac{100}{102} \eta^* \quad (72)$$

Similarly, one can show that $\frac{102}{100} \eta^* \geq \eta$. \square

Lemma 4. Let $U \in \mathbb{C}^{d \times r}$, $U_- \in \mathbb{C}^{d \times r}$, and $U^* \in \mathbb{C}^{d \times r}$, such that $\min_{R \in \mathcal{O}} \|U - U^* R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}}$ and $\min_{R \in \mathcal{O}} \|U_- - U^* R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}}$, where $\rho^* = U^* U^{\star\top}$, and $\kappa := \frac{1 + \delta_{2r}}{1 - \delta_{2r}} > 1$, for $\delta_{2r} \leq \frac{1}{10}$, and $\tau(\rho^*) := \frac{\sigma_1(\rho^*)}{\sigma_r(\rho^*)} > 1$. By Lemma 2, the above imply also that: $\|Z - U^* R_Z\|_F \leq \left(\frac{3}{2} + 2|\mu| \right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}}$. Then, under RIP assumptions of the mapping \mathcal{A} , we have:

$$\begin{aligned} & \left\langle \mathcal{A}^\dagger (\mathcal{A}(Z Z^\dagger) - y), (Z - U^* R_Z)(Z - U^* R_Z)^\dagger \right\rangle \\ & \geq - \left(\theta \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 + \frac{10.1}{100} \beta^2 \cdot \hat{\eta} \cdot \frac{(1+2|\mu|)^2}{\left(1 - (1+2|\mu|) \frac{1}{200}\right)^2} \cdot \left\| \mathcal{A}^\dagger (\mathcal{A}(Z Z^\dagger) - y) \cdot Z \right\|_F^2 \right) \end{aligned} \quad (73)$$

where

$$\theta = \frac{(1 - \delta_{2r}) \left(1 + (1 + 2|\mu|) \frac{1}{200}\right)^2}{10^3} + (1 + \delta_{2r}) \left(2 + (1 + 2|\mu|) \cdot \frac{1}{200}\right) (1 + 2|\mu|) \cdot \frac{1}{200},$$

$$\text{and } \hat{\eta} = \frac{1}{4((1 + \delta_r) \|Z Z^\dagger\|_2 + \|\mathcal{A}^\dagger (\mathcal{A}(Z Z^\dagger) - y) Q_Z Q_Z^\dagger\|_2)}.$$

Proof. First, denote $\Delta := Z - U^* R_Z$. Then:

$$\begin{aligned}
& \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), (Z - U^* R_Z)(Z - U^* R_Z)^\dagger \right\rangle \\
& \stackrel{(i)}{=} \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_\Delta Q_\Delta^\dagger, \Delta_Z \Delta_Z^\dagger \right\rangle \\
& \geq - \left| \text{Tr} \left(\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_\Delta Q_\Delta^\dagger \cdot \Delta_Z \Delta_Z^\dagger \right) \right| \\
& \stackrel{(ii)}{\geq} - \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_\Delta Q_\Delta^\dagger\|_2 \cdot \text{Tr}(\Delta_Z \Delta_Z^\dagger) \\
& \stackrel{(iii)}{\geq} - \left(\|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 + \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_{U^*} Q_{U^*}^\dagger\|_2 \right) \|Z - U^* R_Z\|_F^2
\end{aligned} \tag{74}$$

Note that (i) follows from the fact $\Delta_Z = \Delta_Z Q_\Delta Q_\Delta^\dagger$, for a matrix Q that spans the row space of Δ_Z , and (ii) follows from $|\text{Tr}(AB)| \leq \|A\|_2 \text{Tr}(B)$, for PSD matrix B (Von Neumann's trace inequality [62]). For the transformation in (iii), we use that fact that the row space of Δ_Z , $\text{SPAN}(\Delta_Z)$, is a subset of $\text{SPAN}(Z \cup U^*)$, as Δ_Z is a linear combination of U and U^* .

To bound the first term in equation (74), we observe:

$$\begin{aligned}
& \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(i)}{=} \hat{\eta} \cdot 4 \left((1 + \delta_{2r}) \|ZZ^\dagger\|_2 \right. \\
& \quad \left. + \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \right) \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2 \\
& = \underbrace{4\hat{\eta}(1 + \delta_{2r})\|ZZ^\dagger\|_2 \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2}_{:=A} \\
& \quad + 4\hat{\eta} \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \|Z - U^* R_Z\|_F^2
\end{aligned}$$

where (i) is due to the definition of $\hat{\eta}$.

To bound term A , we observe that $\|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \leq \frac{(1 - \delta_{2r})\sigma_r(ZZ^\dagger)}{10^3}$ or $\|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \geq \frac{(1 - \delta_{2r})\sigma_r(ZZ^\dagger)}{10^3}$. This results into bounding A as follows:

$$4\hat{\eta}(1 + \delta_{2r})\|ZZ^\dagger\|_2 \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2 \tag{75}$$

$$\leq \max \left\{ \frac{4 \cdot \hat{\eta} \cdot (1 + \delta_{2r}) \|ZZ^\dagger\|_2 \cdot (1 - \delta_{2r}) \sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2, \right. \tag{76}$$

$$\left. \hat{\eta} \cdot 4 \cdot 10^3 \kappa_T(ZZ^\dagger) \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \|Z - U^* R_Z\|_F^2 \right\} \tag{77}$$

$$\leq \frac{4 \cdot \hat{\eta} \cdot (1 - \delta_{2r}^2) \|ZZ^\dagger\|_2 \cdot \sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \tag{78}$$

$$+ \hat{\eta} \cdot 4 \cdot 10^3 \kappa_T(ZZ^\dagger) \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \|Z - U^* R_Z\|_F^2. \tag{79}$$

Combining the above inequalities, we obtain:

$$\begin{aligned}
& \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(i)}{\leq} \frac{(1-\delta_{2r})\sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \\
& \quad + (10^3 \kappa \tau(ZZ^\dagger) + 1) \cdot 4 \cdot \hat{\eta} \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(ii)}{\leq} \frac{(1-\delta_{2r})\sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \\
& \quad + (10^3 \beta^2 \kappa \tau(\rho^*) + 1) \cdot 4 \cdot \hat{\eta} \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\kappa \tau(\rho^*)} \cdot \frac{1}{10^6} \sigma_r(\rho^*) \\
& \stackrel{(iii)}{\leq} \frac{(1-\delta_{2r})\sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \\
& \quad + 4 \cdot 1001 \beta^2 \cdot \hat{\eta} \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{10^6 \left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2} \sigma_r(ZZ^\dagger) \\
& \stackrel{(iv)}{\leq} \frac{(1-\delta_{2r})\sigma_r(ZZ^\dagger)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \\
& \quad + 4 \cdot 1001 \beta^2 \cdot \hat{\eta} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{10^6 \left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \\
& \stackrel{(v)}{\leq} \frac{(1-\delta_{2r}) \left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2 \sigma_r(\rho^*)}{10^3} \cdot \|Z - U^* R_Z\|_F^2 \\
& \quad + \frac{1}{200} \beta^2 \cdot \hat{\eta} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2
\end{aligned}$$

where (i) follows from $\hat{\eta} \leq \frac{1}{4(1+\delta_{2r})\|ZZ^\dagger\|_2}$, (ii) is due to Corollary 3, bounding $\|Z - U^* R_Z\|_F \leq \rho \sigma_r(\rho^*)^{1/2}$, where $\rho := \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3 \sqrt{\kappa \tau(\rho^*)}}$ by Lemma 2, (iii) is due to $(10^3 \beta^2 \kappa \tau(\rho^*) + 1) \leq 1001 \beta^2 \kappa \tau(\rho^*)$, and by Corollary 1, (iv) is due to the fact $\sigma_r(ZZ^\dagger) \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger\|_2^2 \leq \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2$, and (v) is due to Corollary 1.

Next, we bound the second term in equation (74):

$$\begin{aligned}
& \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_{U^*} Q_{U^*}^\dagger\|_2 \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(i)}{\leq} \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) - \mathcal{A}^\dagger(\mathcal{A}(\rho^*) - y)\|_2 \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(ii)}{\leq} (1 + \delta_{2r}) \cdot \|ZZ^\dagger - U^* U^{*\dagger}\|_F \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(iii)}{\leq} (1 + \delta_{2r})(2 + \rho) \cdot \rho \cdot \sigma_1(U^*) \cdot \sigma_r(U^*) \cdot \|Z - U^* R_Z\|_F^2 \\
& \stackrel{(iv)}{\leq} (1 + \delta_{2r})(2 + \rho) \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 \\
& \leq (1 + \delta_{2r}) \left(2 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}\right) \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2,
\end{aligned}$$

where (i) follows from $\|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_{U^*} Q_{U^*}^\dagger\|_2 \leq \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)\|_2$ and $\mathcal{A}^\dagger(\mathcal{A}(\rho^*) - y) = 0$, (ii) is due to smoothness of f and the RIP constants, (iii) follows from [12, Lemma 18], for $\rho = \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3 \sqrt{\kappa \tau(\rho^*)}}$, (iv) follows from substituting ρ above, and observing that $\tau(\rho^*) = \sigma_1(U^*)^2 / \sigma_r(U^*)^2 > 1$ and $\kappa = (1 + \delta_{2r}) / (1 - \delta_{2r}) > 1$.

Combining the above we get:

$$\begin{aligned}
& \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), (Z - U^* R_Z)(Z - U^* R_Z)^\dagger \right\rangle \\
& \geq - \left(\theta \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 + \frac{1}{200} \beta^2 \cdot \hat{\eta} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \right) \quad (80)
\end{aligned}$$

where $\theta = \frac{(1-\delta_{2r}) \left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2}{10^3} + (1 + \delta_{2r}) \left(2 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}\right) \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}$. \square

Lemma 5. Under identical assumptions with Lemma 4, the following inequality holds:

$$\left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^*U^{\star\dagger} \right\rangle \geq 1.1172\eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)Z \right\|_F^2 + \frac{1-\delta_{2r}}{2} \|U^*U^{\star\dagger} - ZZ^\dagger\|_F^2 \quad (81)$$

Proof. By smoothness assumption of the objective, based on the RIP assumption, we have:

$$\frac{1}{2} \|\mathcal{A}(ZZ^\dagger) - y\|_2^2 \geq \frac{1}{2} \|\mathcal{A}(U_+U_+^\dagger) - y\|_2^2 \quad (82)$$

$$- \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), U_+U_+^\dagger - ZZ^\dagger \right\rangle - \frac{1+\delta_{2r}}{2} \|U_+U_+^\dagger - ZZ^\dagger\|_F^2 \Rightarrow \quad (83)$$

$$\frac{1}{2} \|\mathcal{A}(ZZ^\dagger) - y\|_2^2 \geq \frac{1}{2} \|\mathcal{A}(U^*U^{\star\dagger}) - y\|_2^2 \quad (84)$$

$$- \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), U_+U_+^\dagger - ZZ^\dagger \right\rangle - \frac{1+\delta_{2r}}{2} \|U_+U_+^\dagger - ZZ^\dagger\|_F^2 \quad (85)$$

due to the optimality $\|\mathcal{A}(U^*U^{\star\dagger}) - y\|_2^2 = 0 \leq \|\mathcal{A}(VV^\dagger) - y\|_2^2$, for any $V \in \mathbb{C}^{d \times r}$. Also, by the restricted strong convexity with RIP, we get:

$$\frac{1}{2} \|\mathcal{A}(U^*U^{\star\dagger}) - y\|_2^2 \geq \frac{1}{2} \|\mathcal{A}(ZZ^\dagger) - y\|_2^2 \quad (86)$$

$$+ \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), U^*U^{\star\dagger} - ZZ^\dagger \right\rangle + \frac{1-\delta_{2r}}{2} \|U^*U^{\star\dagger} - ZZ^\dagger\|_F^2 \quad (87)$$

Adding the two inequalities, we obtain:

$$\left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^*U^{\star\dagger} \right\rangle \geq \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U_+U_+^\dagger \right\rangle \quad (88)$$

$$- \frac{1+\delta_{2r}}{2} \|U_+U_+^\dagger - ZZ^\dagger\|_F^2 + \frac{1-\delta_{2r}}{2} \|U^*U^{\star\dagger} - ZZ^\dagger\|_F^2 \quad (89)$$

To proceed we observe:

$$U_+U_+^\dagger = (Z - \eta\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)Z) \cdot (Z - \eta\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)Z)^\dagger \quad (90)$$

$$= ZZ^\dagger - \eta ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) - \eta\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \quad (91)$$

$$+ \eta^2 \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \quad (92)$$

$$\stackrel{(i)}{=} ZZ^\dagger - \left(I - \frac{\eta}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \right) \cdot \eta ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \quad (93)$$

$$- \eta\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \left(I - \frac{\eta}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \right) \quad (94)$$

where (i) is due to the fact $\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) = \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Q_Z Q_Z^\dagger \cdot ZZ^\dagger \cdot Q_Z Q_Z^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)$, for Q_Z a basis matrix whose columns span the column space of Z ; also, I is the identity matrix whose dimension is apparent from the context. Thus:

$$\frac{\eta}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \preceq \frac{10.5}{10} \frac{\hat{\eta}}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \quad (95)$$

and, hence,

$$I - \frac{\eta}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \succeq I - \frac{10.5}{10} \frac{\hat{\eta}}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y). \quad (96)$$

Define $\Psi = I - \frac{\eta}{2} Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)$. Then, using the definition of $\hat{\eta}$, we know that $\hat{\eta} \leq \frac{1}{4\|Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)\|_2}$, and thus:

$$\Psi \succ 0, \quad \sigma_1(\Psi) \leq 1 + \frac{21}{160}, \quad \text{and} \quad \sigma_n(\Psi) \geq 1 - \frac{21}{160}. \quad (97)$$

Going back to the main recursion and using the above expression for $U_+U_+^\dagger$, we have:

$$\left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^*U^{\star\dagger} \right\rangle - \frac{1-\delta_{2r}}{2} \|U^*U^{\star\dagger} - ZZ^\dagger\|_F^2 \quad (98)$$

$$\geq \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U_+U_+^\dagger \right\rangle - \frac{1+\delta_{2r}}{2} \|U_+U_+^\dagger - ZZ^\dagger\|_F^2 \quad (99)$$

$$\stackrel{(i)}{\geq} 2\eta \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \Psi \right\rangle \quad (100)$$

$$- \frac{1+\delta_{2r}}{2} \|2\eta \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \Psi\|_F^2 \quad (101)$$

$$\stackrel{(ii)}{\geq} 2 \left(1 - \frac{21}{160}\right) \eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \quad (102)$$

$$- 2(1 + \delta_{2r}) \eta^2 \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \cdot \|Z\|_2^2 \cdot \|\Psi\|_2^2 \quad (103)$$

$$\stackrel{(iii)}{\geq} 2 \left(1 - \frac{21}{160}\right) \eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \quad (104)$$

$$- 2(1 + \delta_{2r}) \eta^2 \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \cdot \|Z\|_2^2 \cdot \left(1 + \frac{21}{160}\right)^2 \quad (105)$$

$$= 2 \left(1 - \frac{21}{160}\right) \eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \cdot \left(1 - 2(1 + \delta_{2r}) \eta \cdot \|Z\|_2^2 \cdot \left(1 + \frac{21}{160}\right)^2 \cdot \frac{1}{2(1 - \frac{21}{160})}\right) \quad (106)$$

$$\stackrel{(iv)}{\geq} 2 \left(1 - \frac{21}{160}\right) \eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \cdot \left(1 - 2(1 + \delta_{2r}) \frac{10.5}{10} \hat{\eta} \cdot \|Z\|_2^2 \cdot \left(1 + \frac{21}{160}\right)^2 \cdot \frac{1}{2(1 - \frac{21}{160})}\right) \quad (107)$$

$$\stackrel{(v)}{\geq} 2 \left(1 - \frac{21}{160}\right) \eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \cdot \left(1 - \frac{10.5}{10} \frac{\left(1 + \frac{21}{160}\right)^2}{4(1 - \frac{21}{160})}\right) \quad (108)$$

$$= 1.0656\eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \quad (109)$$

where (i) is due to the symmetry of the objective; (ii) is due to Cauchy-Schwartz inequality and the fact:

$$\left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \Psi \right\rangle \quad (110)$$

$$= \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \right\rangle \quad (111)$$

$$- \frac{\eta}{2} \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \right\rangle \quad (112)$$

$$\stackrel{(i)}{\geq} \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \right\rangle \quad (113)$$

$$- \frac{10.5}{10} \frac{\hat{\eta}}{2} \left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot ZZ^\dagger \cdot \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \right\rangle \quad (114)$$

$$\geq \left(1 - \frac{10.5}{10} \frac{\hat{\eta}}{2} \|Q_Z Q_Z^\dagger \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y)\|_2^2\right) \cdot \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \quad (115)$$

$$\geq \left(1 - \frac{21}{160}\right) \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 \quad (116)$$

where (i) is due to $\eta \leq \frac{10.5}{10} \hat{\eta}$, and the last inequality comes from the definition of the $\hat{\eta}$ and its upper bound; (iii) is due to the upper bound on $\|\Psi\|_2$ above; (iv) is due to $\eta \leq \frac{10.5}{10} \hat{\eta}$; (v) is due to $\hat{\eta} \leq \frac{1}{4(1+\delta_{2r})\|ZZ^\dagger\|_2}$. The above lead to the desiderata:

$$\left\langle \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^*U^{\star\dagger} \right\rangle \geq 1.0656\eta \left\| \mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 + \frac{1-\delta_{2r}}{2} \|U^*U^{\star\dagger} - ZZ^\dagger\|_F^2 \quad (117)$$

□

Detailed proof of Theorem 2

We first denote $U_+ \equiv U_{i+1}$, $U \equiv U_i$, $U_- \equiv U_{i-1}$ and $Z \equiv Z_i$. Let us start with the following equality. For $R_Z \in \mathcal{O}$ as the minimizer of $\min_{R \in \mathcal{O}} \|Z - U^* R\|_F$, we have:

$$\|U_+ - U^* R_Z\|_F^2 = \|U_+ - Z + Z - U^* R_Z\|_F^2 \quad (118)$$

$$= \|U_+ - Z\|_F^2 + \|Z - U^* R_Z\|_F^2 - 2\langle U_+ - Z, U^* R_Z - Z \rangle \quad (119)$$

The proof focuses on how to bound the last part on the right-hand side. By definition of U_+ , we get:

$$\langle U_+ - Z, U^* R_Z - Z \rangle = \langle Z - \eta \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z - Z, U^* R_Z - Z \rangle \quad (120)$$

$$= \eta \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z, Z - U^* R_Z \rangle \quad (121)$$

Observe the following:

$$\langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z, Z - U^* R_Z \rangle = \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^* R_Z Z^\dagger \rangle \quad (122)$$

$$= \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - \frac{1}{2} U^* U^{\dagger} + \frac{1}{2} U^* U^{\dagger} - U^* R_Z Z^\dagger \rangle \quad (123)$$

$$= \frac{1}{2} \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^* U^{\dagger} \rangle \quad (124)$$

$$+ \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), \frac{1}{2} (ZZ^\dagger + U^* U^{\dagger}) - U^* R_Z Z^\dagger \rangle \quad (125)$$

$$= \frac{1}{2} \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^* U^{\dagger} \rangle \quad (126)$$

$$+ \frac{1}{2} \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), (Z - U^* R_Z)(Z - U^* R_Z)^\dagger \rangle \quad (127)$$

By Lemmata 4 and 5, we have:

$$\|U_+ - U^* R_Z\|_F^2 = \|U_+ - Z\|_F^2 + \|Z - U^* R_Z\|_F^2 - 2\langle U_+ - Z, U^* R_Z - Z \rangle \quad (128)$$

$$= \eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2 + \|Z - U^* R_Z\|_F^2 \quad (129)$$

$$- \eta \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), ZZ^\dagger - U^* U^{\dagger} \rangle \quad (130)$$

$$- \eta \langle \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y), (Z - U^* R_Z)(Z - U^* R_Z)^\dagger \rangle \quad (131)$$

$$\leq \eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2 + \|Z - U^* R_Z\|_F^2 \quad (132)$$

$$- 1.0656\eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2 - \eta \frac{1-\delta_{2F}}{2} \|U^* U^{\dagger} - ZZ^\dagger\|_F^2 \quad (133)$$

$$+ \eta \left(\theta \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 \right. \quad (134)$$

$$\left. + \frac{1}{200} \beta^2 \cdot \hat{\eta} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \right) \quad (135)$$

Next, we use the following lemma:

Lemma 6. [88, Lemma 5.4] For any $W, V \in \mathbb{C}^{d \times r}$, the following holds:

$$\|WW^\dagger - VV^\dagger\|_F^2 \geq 2(\sqrt{2} - 1) \cdot \sigma_r(VV^\dagger) \cdot \min_{R \in \mathcal{O}} \|W - VR\|_F^2. \quad (136)$$

From Lemma 6, the quantity $\|U^* U^{\dagger} - ZZ^\dagger\|_F^2$ satisfies:

$$\|U^* U^{\dagger} - ZZ^\dagger\|_F^2 \geq 2(\sqrt{2} - 1) \cdot \sigma_r(\rho^*) \cdot \min_{R \in \mathcal{O}} \|Z - U^* R\|_F^2 = 2(\sqrt{2} - 1) \cdot \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2, \quad (137)$$

which, in our main recursion, results in:

$$\|U_+ - U^* R_Z\|_F^2 \leq \eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2 + \|Z - U^* R_Z\|_F^2 \quad (138)$$

$$- 1.0656\eta^2 \left\| \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 - \eta(\sqrt{2} - 1)(1 - \delta_{2r})\sigma_r(\rho^*) \|Z - U^* R_Z\|_F^2 \quad (139)$$

$$+ \eta \left(\theta \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 \right. \quad (140)$$

$$\left. + \frac{1}{200}\beta^2 \cdot \hat{\eta} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right)\frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \right) \quad (141)$$

$$\stackrel{(i)}{\leq} \eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z\|_F^2 + \|Z - U^* R_Z\|_F^2 \quad (142)$$

$$- 1.0656\eta^2 \left\| \mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) Z \right\|_F^2 - \eta(\sqrt{2} - 1)(1 - \delta_{2r})\sigma_r(\rho^*) \|Z - U^* R_Z\|_F^2 \quad (143)$$

$$+ \eta \left(\theta \sigma_r(\rho^*) \cdot \|Z - U^* R_Z\|_F^2 \right. \quad (144)$$

$$\left. + \frac{1}{200}\beta^2 \cdot \frac{10}{9} \eta \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right)\frac{1}{10^3}\right)^2} \cdot \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \right) \quad (145)$$

$$\stackrel{(ii)}{=} \left(1 + \frac{1}{200}\beta^2 \cdot \frac{10}{9} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right)\frac{1}{10^3}\right)^2} - 1.0656 \right) \eta^2 \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \quad (146)$$

$$+ \left(1 + \eta \theta \sigma_r(\rho^*) - \eta(\sqrt{2} - 1)(1 - \delta_{2r})\sigma_r(\rho^*) \right) \|Z - U^* R_Z\|_F^2 \quad (147)$$

where (i) is due to Lemma 3, and (ii) is due to the definition of U_+ .

Under the facts that $\mu = \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa \tau(\rho^*)}} \cdot \frac{\varepsilon}{4 \cdot \sigma_1(\rho^*)^{1/2} \cdot r}$, for $\varepsilon \in (0, 1)$ user-defined, and $\delta_{2r} \leq \frac{1}{10}$, the main constant quantities in our proof so far simplify into:

$$\beta = \frac{1 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}}{1 - \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3}} = 1.003, \quad \text{and} \quad \beta^2 = 1.006, \quad (148)$$

by Corollary 3. Thus:

$$1 + \frac{1}{200}\beta^2 \cdot \frac{10}{9} \cdot \frac{\left(\frac{3}{2} + 2|\mu|\right)^2}{\left(1 - \left(\frac{3}{2} + 2|\mu|\right)\frac{1}{10^3}\right)^2} - 1.0656 \leq -0.0516, \quad (149)$$

and our recursion becomes:

$$\|U_+ - U^* R_Z\|_F^2 \leq -0.0516 \cdot \eta^2 \cdot \|\mathcal{A}^\dagger (\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \quad (150)$$

$$+ \left(1 + \eta \theta \sigma_r(\rho^*) - \eta(\sqrt{2} - 1)(1 - \delta_{2r})\sigma_r(\rho^*) \right) \|Z - U^* R_Z\|_F^2 \quad (151)$$

Finally,

$$\theta = \frac{(1 - \delta_{2r}) \left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3} \right)^2}{10^3} + (1 + \delta_{2r}) \left(2 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \right) \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \quad (152)$$

$$\stackrel{(i)}{=} (1 - \delta_{2r}) \cdot \left(\frac{\left(1 + \left(\frac{3}{2} + 2|\mu|\right) \frac{1}{10^3} \right)^2}{10^3} + \kappa \left(2 + \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \right) \left(\frac{3}{2} + 2|\mu|\right) \cdot \frac{1}{10^3} \right) \quad (153)$$

$$\leq 0.0047 \cdot (1 - \delta_{2r}). \quad (154)$$

where (i) is by the definition of $\kappa := \frac{1 + \delta_{2r}}{1 - \delta_{2r}} \leq 1.223$ for $\delta_{2r} \leq \frac{1}{10}$, by assumption. Combining the above in

our main inequality, we obtain:

$$\begin{aligned} \|U_+ - U^* R_Z\|_F^2 &\leq -0.0516 \cdot \eta^2 \cdot \|\mathcal{A}^\dagger(\mathcal{A}(ZZ^\dagger) - y) \cdot Z\|_F^2 \\ &\quad + \left(1 + \eta\sigma_r(\rho^*)(1 - \delta_{2r}) \cdot (0.0047 - \sqrt{2} + 1)\right) \|Z - U^* R_Z\|_F^2 \\ &\leq \left(1 - \frac{4\eta\sigma_r(\rho^*)(1 - \delta_{2r})}{10}\right) \|Z - U^* R_Z\|_F^2 \end{aligned} \quad (155)$$

By Lemma 3, we know that $\eta \geq \frac{100}{102}\eta^*$. Also, $\eta^* = \frac{1}{4(1 + \delta_{2r})\|\rho^*\|_2}$, since $\|\mathcal{A}^\dagger(\mathcal{A}(\rho^*) - y)\|_2 = 0$, in the noiseless setting. Returning to (155), we have:

$$\|U_+ - U^* R_Z\|_F^2 \leq \left(1 - 0.393 \cdot \frac{(1 - \delta_{2r})\sigma_r(\rho^*)}{(1 + \delta_{2r})\sigma_1(\rho^*)}\right) \|Z - U^* R_Z\|_F^2 \quad (156)$$

$$= \left(1 - \frac{0.393}{\kappa\tau(\rho^*)}\right) \|Z - U^* R_Z\|_F^2 \quad (157)$$

Taking square root on both sides, we obtain:

$$\|U_+ - U^* R_Z\|_F \leq \sqrt{1 - \frac{0.393}{\kappa\tau(\rho^*)}} \|Z - U^* R_Z\|_F \quad (158)$$

Let us define $\xi = \sqrt{1 - \frac{0.393}{\kappa\tau(\rho^*)}} < 1$. Using the definitions $Z = U + \mu(U - U_-)$ and $R_Z \in \arg \min_{R \in \mathcal{O}} \|Z - U^* R\|_F$, we get

$$\begin{aligned} \|U_+ - U^* R_Z\|_F &\leq \xi \cdot \min_{R \in \mathcal{O}} \|Z - U^* R\|_F = \xi \cdot \min_{R \in \mathcal{O}} \|U + \mu(U - U_-) - U^* R\|_F \\ &= \xi \cdot \min_{R \in \mathcal{O}} \|U + \mu(U - U_-) - (1 - \mu + \mu)U^* R\|_F \\ &\stackrel{(i)}{\leq} \xi \cdot |1 + \mu| \cdot \min_{R \in \mathcal{O}} \|U - U^* R\|_F + \xi \cdot |\mu| \cdot \min_{R \in \mathcal{O}} \|U_- - U^* R\|_F + \xi \cdot |\mu| \cdot r\sigma_1(\rho^*)^{1/2} \end{aligned}$$

where (i) follows from steps similar to those in Lemma 2. Further observe that $\min_{R \in \mathcal{O}} \|U_+ - U^* R\|_F \leq \|U_+ - U^* R_Z\|_F$, thus leading to:

$$\min_{R \in \mathcal{O}} \|U_+ - U^* R\|_F \leq \xi \cdot |1 + \mu| \cdot \min_{R \in \mathcal{O}} \|U - U^* R\|_F + \xi \cdot |\mu| \cdot \min_{R \in \mathcal{O}} \|U_- - U^* R\|_F + \xi \cdot |\mu| \cdot r\sigma_1(\rho^*)^{1/2} \quad (159)$$

Including two subsequent iterations in a single two-dimensional first-order system, we get the following characterization:

$$\begin{bmatrix} \min_{R \in \mathcal{O}} \|U_{i+1} - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_i - U^* R\|_F \end{bmatrix} \leq \underbrace{\begin{bmatrix} \xi \cdot |1 + \mu| & \xi \cdot |\mu| \\ 1 & 0 \end{bmatrix}}_{:=A} \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_i - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{i-1} - U^* R\|_F \end{bmatrix} \quad (160)$$

$$+ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r, \quad (161)$$

Observe that the contraction matrix A has non-negative values. Unfolding the above recursion for $J + 1$ iterations, we obtain:

$$\begin{bmatrix} \min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_J - U^* R\|_F \end{bmatrix} \leq A^{J+1} \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} \quad (162)$$

$$+ \left(\sum_{i=0}^J A^i\right) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (163)$$

Let us focus on the properties of the matrix A . We re-use Lemma 2 in [49], after appropriate changes:

Lemma 7. Let A be the 2×2 matrix, as defined above, parameterized by $0 < \xi = \sqrt{1 - \frac{0.393}{\kappa\tau(\rho^*)}} < 1$, and user-defined parameter μ . The characteristic polynomial of A is defined as:

$$\lambda^2 - \text{Tr}(A) \cdot \lambda + \det(A) = 0 \quad (164)$$

where λ represent the eigenvalue(s) of A . Define $\Delta := \text{Tr}(A)^2 - 4 \cdot \det(A) = \xi^2 \cdot (1 + \mu)^2 + 4\xi \cdot |\mu|$. Then, the eigenvalues of A satisfy the expression: $\lambda_{1,2} = \frac{\xi \cdot |1 + \mu| \pm \sqrt{\Delta}}{2}$.

A proof of Lemma 7 follows immediately from the quadratic formula. In our case, $\Delta > 0$, which means that

$$\lambda_{1,2} = \frac{\sqrt{1 - \frac{0.393}{\kappa\tau(\rho^*)}} \cdot |1 + \mu| \pm \sqrt{\left(1 - \frac{0.393}{\kappa\tau(\rho^*)}\right) (1 + \mu)^2 + 4\sqrt{1 - \frac{0.393}{\kappa\tau(\rho^*)}} |\mu|}}{2}. \quad (165)$$

The following lemma describes how one can compute a power of a 2×2 matrix A , A^i , through the eigenvalues $\lambda_{1,2}$ (real and distinct eigenvalues); the proof is provided in [49].

Lemma 8 ([93]). *Let A be a 2×2 matrix with real eigenvalues $\lambda_{1,2}$. Then, the following expression holds, when $\lambda_1 \neq \lambda_2$:*

$$A^i = \frac{\lambda_1^i - \lambda_2^i}{\lambda_1 - \lambda_2} \cdot A - \lambda_1 \lambda_2 \cdot \frac{\lambda_1^{i-1} - \lambda_2^{i-1}}{\lambda_1 - \lambda_2} \cdot I \quad (166)$$

where λ_i denotes the i -th eigenvalue of A in order.

For the rest of the proof, we will require $|\lambda_i(A)| < 1$; we will discuss later in the text what this requirement means in terms of the parameters of the problem. Under this requirement, $\sum_{i=0}^J A^i$ converges to: $\sum_{i=0}^J A^i = (I - A)^{-1} (I - A^{J+1})$, where:

$$B := (I - A)^{-1} \stackrel{|\mu| \ll \mu}{=} \frac{1}{1 - \xi(1 + 2\mu)} \cdot \begin{bmatrix} 1 & \xi\mu \\ 1 & 1 - \xi(1 + \mu) \end{bmatrix} \quad (167)$$

This transforms our recursion into:

$$\begin{bmatrix} \min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_J - U^* R\|_F \end{bmatrix} \quad (168)$$

$$\leq A^{J+1} \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} - B \cdot A^{J+1} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (169)$$

$$+ B \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (170)$$

$$\stackrel{(i)}{\leq} \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot A \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} + |\lambda_1| \cdot \frac{2|\lambda_1|^J}{|\lambda_1| - |\lambda_2|} \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} \quad (171)$$

$$+ \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot B \cdot A \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r + |\lambda_1| \cdot \frac{2|\lambda_1|^J}{|\lambda_1| - |\lambda_2|} \cdot B \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (172)$$

$$+ B \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (173)$$

$$\stackrel{(ii)}{\leq} \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot (A + I) \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} + \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot B \cdot (A + I) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (174)$$

$$+ B \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (175)$$

$$\stackrel{(iii)}{\leq} \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \begin{bmatrix} 1 + \xi \cdot |1 + \mu| & \xi \cdot |\mu| \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \\ \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F \end{bmatrix} \quad (176)$$

$$+ \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \frac{1}{1 - \xi(1 + 2\mu)} \cdot \begin{bmatrix} 1 + \xi(1 + 2\mu) \\ 2 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (177)$$

$$+ \frac{1}{1 - \xi(1 + 2\mu)} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \xi \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (178)$$

where for (i) we used Lemma 8, and the fact that B and A have non-negative values to retain the bound, (ii) is due to $|\lambda_1| < 1$, (iii) is obtained by explicitly computing $_{BA} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $_{B} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

Focusing on the top term of this expression and under the assumption that $\min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F = \min_{R \in \mathcal{O}} \|U_{-1} - U^* R\|_F$, we have:

$$\min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \leq \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \left((1 + \xi(1 + 2\mu)) \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F + \frac{\xi(1 + \xi(1 + 2\mu))}{1 - \xi(1 + 2\mu)} \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \right) \quad (179)$$

$$+ \frac{\xi\mu}{1 - \xi(1 + 2\mu)} \cdot \sigma_1(\rho^*)^{1/2} \cdot r \quad (180)$$

We will use the assumption on μ ; then

$$\frac{\xi(1+\xi(1+2\mu))}{1-\xi(1+2\mu)} \cdot |\mu| \cdot \sigma_1(\rho^*)^{1/2} \cdot r \leq \frac{\xi(1+\xi(1+\frac{1}{10^3}))}{1-\xi(1+\frac{1}{10^3})} \cdot \frac{\varepsilon \cdot \sigma_r(\rho^*)^{1/2}}{4 \cdot 10^3 \sqrt{\kappa\tau(\rho^*)}}. \quad (181)$$

and

$$\frac{\xi|\mu|}{1-\xi(1+2\mu)} \cdot \sigma_1(\rho^*)^{1/2} \cdot r \leq \frac{\xi}{1-\xi(1+\frac{1}{10^3})} \cdot \frac{\varepsilon \cdot \sigma_r(\rho^*)^{1/2}}{4 \cdot 10^3 \sqrt{\kappa\tau(\rho^*)}} \quad (182)$$

which further leads to:

$$\min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \leq \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \left((1 + \xi(1 + 2\mu)) \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F + \frac{\xi(1+\xi(1+\frac{1}{10^3}))}{1-\xi(1+\frac{1}{10^3})} \cdot \frac{\varepsilon \cdot \sigma_r(\rho^*)^{1/2}}{4 \cdot 10^3 \sqrt{\kappa\tau(\rho^*)}} \right) \quad (183)$$

$$+ \frac{\xi}{1-\xi(1+\frac{1}{10^3})} \cdot \frac{\varepsilon \cdot \sigma_r(\rho^*)^{1/2}}{4 \cdot 10^3 \sqrt{\kappa\tau(\rho^*)}} \quad (184)$$

Under the assumption that $\min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}}$, we get the expression:

$$\min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \leq \frac{2|\lambda_1|^{J+1}}{|\lambda_1| - |\lambda_2|} \cdot \left((1 + \xi(1 + 2\mu)) + \frac{\xi(1+\xi(1+\frac{1}{10^3}))}{1-\xi(1+\frac{1}{10^3})} \cdot \frac{\varepsilon}{4} \right) \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \quad (185)$$

$$+ \frac{\xi \cdot \varepsilon}{4(1-\xi(1+\frac{1}{10^3}))} \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3 \sqrt{\kappa\tau(\rho^*)}} \quad (186)$$

This is the rigorous contraction iteration for MiFGD. A more interpretable description of the above could be:

$$\min_{R \in \mathcal{O}} \|U_{J+1} - U^* R\|_F \lesssim c^{J+1} \cdot \min_{R \in \mathcal{O}} \|U_0 - U^* R\|_F + O(\mu),$$

for a constant $c < 1$. In words, the proposed algorithm achieves a linear convergence rate in iterate distances (first term on RHS), up to a constant proportional to the momentum hyperparameter μ (second term on RHS).

The above hold assuming $|\lambda_i(A)| < 1$; we now focus on the requirement that $|\lambda_i(A)| < 1$. Based on the expression in (165), and using the facts that: *i*) for $\delta_{2r} \leq \frac{1}{10}$, we have $\kappa \leq 1.223$, and *ii*) by assumption $\mu \leq \frac{1}{10^3}$, we observe that $\lambda_i(A) < 1$ for $\tau(\rho^*) \leq 78$; see also Figure 33. For our theory next, we will assume

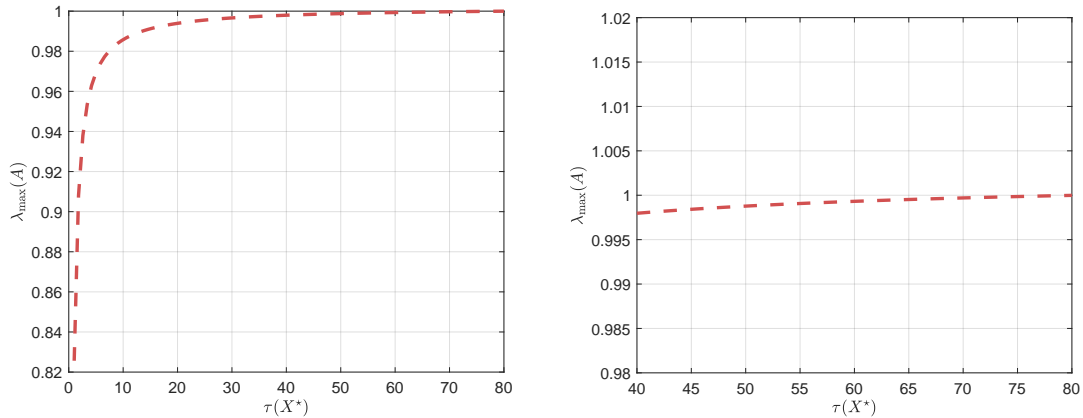


Figure 33: Behavior of maximum $|\lambda_i(A)|$ for various values of $\tau(\rho^*)$; right plot is a zoomed-in version of the left plot.

$\tau(\rho^*) \leq 50$, so that the following requirement holds: In (159), we observe that if $\min_{R \in \mathcal{O}} \|U - U^* R\|_F \leq$

$\frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}}$ and $\min_{R \in \mathcal{O}} \|U_- - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}}$, and under the assumptions on μ , we have:

$$\min_{R \in \mathcal{O}} \|U_+ - U^*R\|_F \leq \xi \cdot |1 + \mu| \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}} + \xi \cdot |\mu| \cdot \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}} + \xi|\mu| \cdot r\sigma_1(\rho^*)^{1/2} \quad (187)$$

$$\leq \xi \left(1 + \frac{3}{10^3}\right) \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}} \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}} \quad (188)$$

for $\xi < 0.9968$ for $\tau(\rho^*) \leq 50$, and for any $\varepsilon \in (0, 1)$. *I.e.*, U_+ satisfies $\min_{R \in \mathcal{O}} \|U_+ - U^*R\|_F \leq \frac{\sigma_r(\rho^*)^{1/2}}{10^3\sqrt{\kappa\tau(\rho^*)}}$. Since the distance remains bounded after each iteration, Lemma 6 hold for all i .