Automatica 114 (2020) 108837

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

# Quantum tomography by regularized linear regressions[☆]

Biqiang Mu [a], Hongsheng Qi [a,b,*], Ian R. Petersen [c], Guodong Shi [d]

[a] Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, PR China
[c] Research School of Electrical, Energy and Materials Engineering, The Australian National University, Canberra 0200, Australia
[d] Australian Center for Field Robotics, School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

*Article history:*
Received 26 April 2019
Received in revised form 30 August 2019
Accepted 25 November 2019
Available online xxxx

*Keywords:*
Quantum state tomography
Linear regression
Regularization

ABSTRACT

In this paper, we study extended linear regression approaches for quantum state tomography based on regularization techniques. For unknown quantum states represented by density matrices, performing measurements under certain basis yields random outcomes, from which a classical linear regression model can be established. First of all, for complete or over-complete measurement bases, we show that the empirical data can be utilized for the construction of a weighted least squares estimate (LSE) for quantum tomography. Taking into consideration the trace-one condition, a constrained weighted LSE can be explicitly computed, being the optimal unbiased estimation among all linear estimators. Next, for general measurement bases, we show that $\ell_2$-regularization with proper regularization gain provides even a lower mean-square error under a cost in bias. The optimal regularization parameter is defined in terms of a risk characterization for any finite sample size and a resulting implementable estimator is proposed. Finally, a concise and unified formula is established for the regularization parameter with complete measurement basis under an equivalent regression model, which proves that the proposed implementable tuning estimator is asymptotically optimal as the number of copies grows to infinity. Additionally, several numerical examples are provided to validate the established results.



## 1. Introduction

Since Feynman pointed out the possibility of using quantum resources to carry out computation in the early 1980s, significant progresses have been made in both the theoretical understanding and the real-world implementations for computing and communication mechanisms based on quantum states (Nielsen & Chuang, 2001). Underpinning such efforts lies in the development of quantum tomography (Artiles, Gill, & Guta, 2005; James, Kwiat, Munro, & White, 2001; Senko et al., 2014; Wootters & Fields, 1989), where reliable quantum state estimation (Bisio, Chiribella, D'Ariano, Facchini, & Perinotti, 2009; Blume-Kohout, 2010; Teo, Zhu, Englert, Rehacek, & Hradil, 2011) and system identification methods (Bonnabel, Mirrahimi, & Rouchon, 2009; Leghtas, Turinici, Rabitz, & Rouchon, 2012; Wang et al., 2018, 2019; Xue, Zhang, & Petersen, 2019) provide basic assurance for the validity of quantum systems that we intend to work on. The fundamental quantum measurement postulate indicates that any form of quantum information probe would have an inherent probabilistic nature. The exponentially growing complexity of quantum systems along with the increasing scale further adds to the challenging reality: only partial information can be made available via measurements for uncertain quantum systems; processing the measurement data faces enormously high computation barrier for large-scale quantum systems.

One primary task of quantum tomography is to determine an unknown quantum state from a number of identical copies (Bisio et al., 2009; Blume-Kohout, 2010; Teo et al., 2011). Performing measurement on those copies along certain observables, i.e., measurement bases, yields independent realizations of some hidden random variable whose statistics encode the quantum state and the observable. Therefore, utilizing the outcomes of the measurements we can build estimations of the unknown quantum state since the observables are known (which can be selected and designed). Apparently the choice of the estimation method is not

☆ This research was supported in part by the National Key R&D Program of China under Grant 2018YFA0703800, the National Natural Science Foundation of China under Grant 61873262, and the Australian Research Council under Grants DP180101805 and DP190103615. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Er-Wei Bai under the direction of Editor Torsten Söderström.

* Corresponding author at: Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

*E-mail addresses:* bqmu@amss.ac.cn (B. Mu), qihongsh@amss.ac.cn (H. Qi), ian.petersen@anu.edu.au (I.R. Petersen), guodong.shi@sydney.edu.au (G. Shi).

https://doi.org/10.1016/j.automatica.2020.108837
0005-1098/

unique since the estimation error metrics can be characterized by different metrics, and the resulting computational feasibility also raises constraint on the potentially viable estimation approaches. Therefore, there is often a tradeoff between estimation quality and computational efficiency (Smolin, Gambetta, & Smith, 2012).

Linear regression, as a universal estimator (Ljung, 1999), becomes a natural and important quantum state tomography approach due to its simplicity and practicability. A thorough comparison was made in Qi et al. (2013) between linear regression method and maximum likelihood estimation for quantum state tomography under complete or over-complete measurement bases. Linear regression was also applied to quantum tomography with incomplete measurement bases for low-rank or sparse quantum states (Alquier, Butucea, Hebiri, & Meziani, 2013; Cai, Kim, Wang, Yuan, & Zhou, 2016; Gross, 2011; Gross, Liu, Flammia, & S. Becker, 2010). In particular, in Gross et al. (2010), quantum state reconstruction was considered in view of the insights from compressed sensing. In Cai et al. (2016), the minimax optimal rates for the quantum state tomography were established when assuming that the quantum state is approximately sparse under the Pauli basis. Recently, linear regression method was also generalized to the adaptive measurement case where selection of the measurement basis depends on the previous measurement outcomes (Qi, Hou, Wang, Dong, Zhong, Li, et al., 2017).

In this paper, we study the role of regularization for linear regression-based quantum state tomography. In recent years, the power of regularization has been well noted in the literature of optimization, machine learning, and system identification, e.g., Chiuso (2016), Goodfellow, Bengio, and Courville (2016) and Shalev-Shwartz (2012) for the purpose of avoiding overfitting in empirical learning. Noting any quantum state can be represented as a trace-one positive Hermitian density matrix, which is of low rank if it is a combination of a small number of pure states, we establish the following results.

- For complete or over-complete measurement basis, the empirical data can be utilized for constructing of a weighted least squares estimate (LSE) for quantum tomography. The weighted LSE provides reduced mean-square error compared to standard LSE. Taking into consideration the trace-one condition, the constrained weighted LSE can be explicitly computed, which is the optimal unbiased estimation that is affine in the measurement data.
- For any (complete, over-complete, or under-complete) measurement basis, a closed form solution is established for tomography with $\ell_2$-regularized weighted linear regression. It is shown that with proper regularization parameter, this regularized regression always provides even a lower mean-square error subject to, of course, a price of additional bias.
- The optimal regularization parameter is characterized in terms of a risk characterization for any finite sample size. An explicit formula is established for the regularization parameter under an equivalent regression model, which proves that the proposed implementable tuning estimator is asymptotically optimal for complete bases as the number of copies grows to infinity for the risk metric.

Numerical examples are provided for the validation of the established theoretical results, which confirms the potential usefulness of the proposed linear regression methods in quantum state tomography.

The remainder of the paper is organized as follows. In Section 2, we present the standard linear regression model for quantum state tomography, and review some preliminary knowledge on the underlying rationale. In Section 3, we present the quantum tomography methods based on LSE, weighted LSE, constrained weighted LSE, and constrained regularized weighted LSE, respectively, whose performances in terms of mean-square error are thoroughly investigated. Section 4 further presents the asymptotically optimal regularization gain under an equivalent model. Numerical examples are presented in Section 5, and finally, some concluding remarks are given in Section 6.

## 2. Problem definition and preliminaries

### 2.1. Linear regression for quantum state tomography

Let $\mathcal{H}$ be a Hilbert space with dimension $d$ that characterizes the state space of a quantum system. Denote the space of linear Hermitian operators over $\mathcal{H}$ by $\mathcal{L}(\mathcal{H})$. Suppose that $\left\{\mathsf{B}_i\right\}_{i=1}^{d^2}$ is an orthonormal basis of $\mathcal{L}(\mathcal{H})$ with $\mathrm{Tr}(\mathsf{B}_i^{\dagger}\mathsf{B}_j) = \delta_{ij}$ and $\mathsf{B}_i^{\dagger} = \mathsf{B}_i$, where $\mathrm{Tr}(\cdot)$ means the trace of a square matrix, $(\cdot)^{\dagger}$ represents the Hermitian conjugate of a complex matrix, and $\delta_{ij}$ is the Kronecker function. A quantum state $\rho$ as a density operator over $\mathcal{H}$ can then be expressed by

$$\rho = \sum_{i=1}^{d^2} \theta_i \mathsf{B}_i \tag{1}$$

where $\theta_i = \mathrm{Tr}(\rho \mathsf{B}_i) \in \mathbb{R}$ is the coordinate of $\rho$ under the given basis $\left\{\mathsf{B}_i\right\}_{i=1}^{d^2}$. Let there be a positive operator-valued measurement (POVM) over the space $\mathcal{H}$, denoted by $\{\mathsf{M}_m\}_{m=1}^{M}$ with $\sum_{m=1}^{M} \mathsf{M}_m^{\dagger}\mathsf{M}_m = \mathsf{I}$, where $\mathsf{I}$ is the identity operator and $M$ is the number of measurement outcomes. Then $\mathsf{E}_m \triangleq \mathsf{M}_m^{\dagger}\mathsf{M}_m$ can be expressed as a linear combination of the orthogonal basis of $\left\{\mathsf{B}_i\right\}_{i=1}^{d^2}$:

$$\mathsf{E}_m = \sum_{i=1}^{d^2} \beta_{mi} \mathsf{B}_i$$

for each $1 \le m \le M$, where $\beta_{mi} = \mathrm{Tr}(\mathsf{E}_m \mathsf{B}_i)$. When the quantum state $\rho$ is being measured under the POVM $\{\mathsf{M}_m\}_{m=1}^{M}$, the probability of observing outcome $m$ is $p_m = \mathrm{Tr}(\mathsf{E}_m\rho) = \boldsymbol{\beta}_m^{\top}\boldsymbol{\theta}$, where $\boldsymbol{\beta}_m = [\beta_{m1}, \ldots, \beta_{md^2}]^{\top}$ and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{d^2}]^{\top}$. Denoting $\mathbf{p} = [p_1, \ldots, p_M]^{\top} \in \mathbb{R}^M$, $\mathbf{A} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M]^{\top} \in \mathbb{R}^{M\times d^2}$, we have the following fundamental quantum measurement description in the form of a linear algebraic equation:

$$\mathbf{p} = \mathbf{A}\boldsymbol{\theta}.$$

The tomography of an unknown quantum state $\rho$ is therefore equivalent to identifying the vector $\boldsymbol{\theta}$, where $\mathbf{A}$ is known and $\mathbf{p}$ is estimated by experimental realizations of measuring $\rho$ from the POVM $\{\mathsf{M}_m\}_{m=1}^{M}$. The POVM can be in general represented under Pauli matrices, see e.g, Cai et al. (2016) and Wang (2013).

A standard quantum state tomography process is as follows: (i) Prepare $N = nM$ identical copies of an uncertain quantum state $\rho$; (ii) Perform measurement along each $\mathsf{M}_m$ within the POVM $\{\mathsf{M}_m\}_{m=1}^{M}$ independently for $n$ copies; (iii) For each $1 \le m \le M$, record the number of times that the measurement operator $\mathsf{M}_m$ is observed among those $n$ experiments, denoted by $\#m$, from the $n$ experiments. Then,

$$\widehat{p}_m = \frac{\#m}{n} \tag{2}$$

is a natural estimator of the probability $p_m$, leading to

$$\widehat{p}_m = \boldsymbol{\beta}_m^{\top}\theta + e_m, \tag{3}$$

where $e_m = \widehat{p}_m - p_m$ is the estimation error. The distribution of $e_m$ depends on the number of copies $n$, as it is the sum of $n$ identical and independently distributed (i.i.d.) Bernoulli random variables with mean $p_m$. This naturally yields the following linear regression problem:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \tag{4}$$

with $\mathbf{y} = [\widehat{p}_1, \ldots, \widehat{p}_M]^\top$ and $\mathbf{e} = [e_1, \ldots, e_M]^\top$.

### 2.2. The noise distribution

Define i.i.d. random variables $b_l^{(m)}$ for $1 \le l \le n$, which takes value 1 with probability $p_m$ and 0 with probability $1 - p_m$. Then there holds

$$e_m = \widehat{p}_m - p_m = \frac{\sum_{l=1}^n b_l^{(m)}}{n} - p_m = \sum_{l=1}^n \frac{b_l^{(m)} - p_m}{n}. \tag{5}$$

Note that $(b_l^{(m)} - p_m)/n$ takes value $(1 - p_m)/n$ with probability $p_m$ and $-p_m/n$ with probability $1 - p_m$. It follows that

$$\mathbb{E}(e_m) = 0 \tag{6a}$$

$$\mathbb{V}(e_m) = \mathbb{E}(e_m)^2 = (p_m - p_m^2)/n. \tag{6b}$$

As a result, the distribution of $e_m$ is as follows:

$$\mathbb{P}\left(e_m = \left(\frac{1-p_m}{n}\right)^k \left(-\frac{p_m}{n}\right)^{n-k}\right)$$
$$= \binom{n}{k} p_m^k (1 - p_m)^{n-k}. \tag{7}$$

As $n$ tends to infinity, each $e_m$ will converge to a Gaussian random variable with mean 0 and variance $(p_m - p_m^2)/n$.

### 2.3. Simultaneous measurements

In the tomography process described above, each $\mathsf{M}_m$ is separately measured, i.e., a binary outcome is recorded for any copy of $\rho$, where 1 represents $\mathsf{M}_m$, and 0 represents $\mathsf{I} - \mathsf{M}_m$. An alternative quantum tomography process can be described based on $n$ copies of $\rho$, where we perform measurement by the POVM $\{\mathsf{M}_m\}_{m=1}^M$ collectively. To be precise, the outcome associated with each copy of the quantum state now takes value in $\{1, \ldots, M\}$, and then the number of times that the outcome $m$ is observed among those $n$ experiments, denoted by $\#'m$, is recorded from the $n$ experiments for each $1 \le m \le M$. For low dimensional quantum systems, e.g., two photon entanglements, such simultaneous measurements are often possible, e.g., Ahnert and Payne (2006). Consequently,

$$\bar{p}_m = \frac{\#'m}{n} \tag{8}$$

is still an estimator of the probability $p_m$, leading to

$$\bar{p}_m = \boldsymbol{\beta}_m^\top \theta + \bar{e}_m, \quad m = 1, \ldots, M. \tag{9}$$

The estimation error $\bar{e}_m$ as a random variable has the same distribution of $e_m$. However, the $\bar{e}_m$ are no longer independent since now $\sum_{m=1}^n \bar{p}_m = 1$ is a sure event. Except for this minor difference, this new formulation of quantum tomography procedure remains the same.

## 3. Regularized linear regressions

In this section, we present a few estimators as quantum tomography solutions for the linear regression (4), and investigate their performances in terms of the mean-square error. The linear regression (4) has the following special features from the underlying quantum tomography procedure:

(i) Heteroscedasticity: the errors $e_m, 1 \le m \le M$ have different variances;
(ii) Trace unity: $\mathrm{Tr}(\rho) = 1$ implies that $\sum_{i=1}^{d^2} \theta_i \mathrm{Tr}(\mathsf{B}_i) = 1$;
(iii) The density operator $\rho$ is often of low rank, e.g., Gross et al. (2010).

In the following, we embed these features successively and derive the corresponding estimators.

### 3.1. Standard least squares

For the estimation problem (4), the least squares (LS) solution (Qi et al., 2013; Rao, Toutenburg, Shalabh, & Heumann, 2008)

$$\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})$$
$$= (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \tag{10}$$

is a common choice provided that $\mathbf{A}$ has full column rank. The estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ admits the following properties:

- $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ is unbiased, namely, $\mathbb{E}\big(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}\big) = \boldsymbol{\theta}$;
- The mean squared error (MSE) matrix of $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ is

$$\mathbb{MSE}\big(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}\big) \triangleq \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta})^\top$$
$$= (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathsf{P} \mathbf{A} (\mathbf{A}^\top \mathbf{A})^{-1} \tag{11}$$

where $\mathsf{P} = \mathrm{diag}\big([p_1 - p_1^2, \ldots, p_M - p_M^2]\big)/n$.

Note that the standard LS estimator neglects the fact that the $e_m$ have different variances, although all of them are zero mean. As a result, the above covariance is not optimal. Furthermore, the condition that $\mathbf{A}$ be full column rank means the POVM $\{\mathsf{M}_m\}_{m=1}^M$ is informationally complete, i.e., any two density operators are distinguishable under the POVM given sufficiently large number of copies. This is not practical for large-scale quantum systems. Meanwhile, the standard LS estimator does not consider the unit trace constraint and often low rank of the quantum state $\rho$.

### 3.2. Weighted regression

Noticing $\mathbb{V}(e_m) = (p_m - p_m^2)/n$, we can instead use the following weighted least squares (WLS) estimate

$$\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top \mathsf{W} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})$$
$$= (\mathbf{A}^\top \mathsf{W} \mathbf{A})^{-1} \mathbf{A}^\top \mathsf{W} \mathbf{y} \tag{12}$$

with $\mathsf{W} = \mathsf{P}^{-1} = n\mathrm{diag}\big([1/(p_1 - p_1^2), \ldots, 1/(p_M - p_M^2)]\big)$ adjusting the difference in variances for the noises $e_m$. This weighted least square $\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}}$ continues to be unbiased since $\mathbb{E}\big(\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}}\big) = \boldsymbol{\theta}$ is easily verifiable and its MSE is

$$\mathbb{MSE}\big(\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}}\big) = (\mathbf{A}^\top \mathsf{W} \mathbf{A})^{-1}. \tag{13}$$

Suppose $\mathrm{rank}(\mathbf{A}) = d^2$ and let $\widehat{\boldsymbol{\theta}}$ be any linear unbiased estimate for $\boldsymbol{\theta}$. Then we have

$$\mathbb{MSE}(\widehat{\boldsymbol{\theta}}) \ge \mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}}).$$

This means it is the best estimator of $\boldsymbol{\theta}$ among all unbiased linear estimators in the sense that it achieves the minimal covariance.

In practice, the matrix $\mathsf{W}$ in (12) is unknown and a feasible solution is to use the estimate

$$\widehat{\boldsymbol{\theta}}^{\mathrm{AWLS}} = (\mathbf{A}^\top \widehat{\mathsf{W}} \mathbf{A})^{-1} \mathbf{A}^\top \widehat{\mathsf{W}} \mathbf{y}, \tag{14}$$

where $\mathsf{W}$ in (12) is replaced by its consistent estimate

$$\widehat{\mathsf{W}} = n \cdot \mathrm{diag}\big([1/(\widehat{p}_1 - \widehat{p}_1^2), \ldots, 1/(\widehat{p}_M - \widehat{p}_M^2)]\big) \tag{15}$$

with $\widehat{p}_m$, $1 \le m \le M$ given by (2). In the following, it is shown that the estimate (14) is accurate enough and asymptotically coincides with (12).

For a random sequence $\xi_n$, we define $\xi_n = O_p(a_n)$ by that $\{\xi_n/a_n\}$ is bounded in probability, i.e., $\forall \epsilon > 0, \exists L > 0$ such that $\mathbb{P}(|\xi_n/a_n| > L) < \epsilon,\ \forall n$. Then there holds for large $n$ that

$$\begin{aligned}
&\widehat{\boldsymbol{\theta}}^{\text{AWLS}} - \widehat{\boldsymbol{\theta}}^{\text{WLS}} \\
&= (\mathbf{A}^\top \widehat{\mathsf{W}}\mathbf{A})^{-1}\mathbf{A}^\top \widehat{\mathsf{W}}\mathbf{e} - (\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\mathbf{A}^\top \mathsf{W}\mathbf{e} \\
&= \left((\mathbf{A}^\top \widehat{\mathsf{W}}\mathbf{A})^{-1} - (\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\right)\mathbf{A}^\top \widehat{\mathsf{W}}\mathbf{e} \\
&\quad + (\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\mathbf{A}^\top\left(\widehat{\mathsf{W}} - \mathsf{W}\right)\mathbf{e} \\
&= O_p(1/\sqrt{n})(\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\mathbf{A}^\top \mathsf{W}\left(1 + O_p(1/\sqrt{n})\right)\mathbf{e} \\
&\quad + (\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\mathbf{A}^\top O_p\left(1/\sqrt{n}\right)\mathsf{W}\mathbf{e} \\
&= O_p(1/\sqrt{n})(\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}\mathbf{A}^\top \mathsf{W}\mathbf{e} \qquad (16)
\end{aligned}$$

in terms of

$$e_m = O_p(1/\sqrt{n}),\ \ 1 \le m \le M$$

and further

$$\widehat{\mathsf{W}} = \mathsf{W}\left(1 + O_p(1/\sqrt{n})\right)$$
$$\widehat{\mathsf{W}} - \mathsf{W} = O_p\left(1/\sqrt{n}\right)\mathsf{W}.$$

This means the difference between $\widehat{\boldsymbol{\theta}}^{\text{AWLS}}$ and $\widehat{\boldsymbol{\theta}}^{\text{WLS}}$ is asymptotically ignorable in comparison with the estimation error of the weighted LSE $\widehat{\boldsymbol{\theta}}^{\text{WLS}}$. Actually, the approximation (15) and resulting conclusions (16) also hold for the following introduced estimators.

### 3.3. Constrained weighted regression

The standard or weighted least squares solutions might lead to estimates that are not legitimate quantum states. In fact, the quantum state has an essential requirement

$$\text{Tr}(\rho) = 1. \qquad (17)$$

This becomes for the model (1) that

$$\boldsymbol{\theta}^\top \text{Tr}(\mathsf{B}) = 1 \qquad (18)$$

where $\text{Tr}(\mathsf{B})$ is defined by

$$\text{Tr}(\mathsf{B}) \triangleq [\text{Tr}(\mathsf{B}_1), \dots, \text{Tr}(\mathsf{B}_{d^2})]^\top. \qquad (19)$$

This inspires us to define the constrained least squares (CLS) estimate

$$\widehat{\boldsymbol{\theta}}^{\text{CLS}} = \mathop{\arg\min}_{\boldsymbol{\theta}^\top \text{Tr}(\mathsf{B})=1}\ (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}). \qquad (20)$$

For the estimate (20), we have the following proposition to characterize its property.

**Proposition 1.** *Suppose* $\text{rank}(\mathbf{A}) = d^2$. *The CLS estimate* $\widehat{\boldsymbol{\theta}}^{\text{CLS}}$ *has the following closed-form solution*

$$\widehat{\boldsymbol{\theta}}^{\text{CLS}} = \widehat{\boldsymbol{\theta}}^{\text{LS}} - \frac{C\text{Tr}(\mathsf{B})}{\text{Tr}(\mathsf{B})^\top C \text{Tr}(\mathsf{B})}\left(\text{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\text{LS}} - 1\right) \qquad (21)$$

*where* $\widehat{\boldsymbol{\theta}}^{\text{LS}}$ *is the least squares estimate given by* (10) *and* $C = (\mathbf{A}^\top \mathbf{A})^{-1}$, *and its MSE matrix is*

$$\mathbb{MSE}\left(\widehat{\boldsymbol{\theta}}^{\text{CLS}}\right) \triangleq \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\text{CLS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\text{CLS}} - \boldsymbol{\theta})^\top = F\mathbf{A}^\top \mathsf{W}^{-1}\mathbf{A}F$$

*where* $F = C - \frac{C\text{Tr}(\mathsf{B})\text{Tr}(\mathsf{B})^\top C}{\text{Tr}(\mathsf{B})^\top C\text{Tr}(\mathsf{B})}$.

To make the notation simple, we will abuse the symbols $F$ and $C$ a little for different cases in the following.

To reduce the MSE of the estimate (20), we can similarly introduce the constrained weighted least squares (CWLS) estimate

$$\widehat{\boldsymbol{\theta}}^{\text{CWLS}} = \mathop{\arg\min}_{\boldsymbol{\theta}^\top \text{Tr}(\mathsf{B})=1}\ (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top \mathsf{W}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}). \qquad (22)$$

**Theorem 1.** *Suppose* $\text{rank}(\mathbf{A}) = d^2$ *and* $p_m \in (0, 1)$ *for* $m = 1, \dots, M$. *The estimate* $\widehat{\boldsymbol{\theta}}^{\text{CWLS}}$ *can be explicitly written as*

$$\widehat{\boldsymbol{\theta}}^{\text{CWLS}} = \widehat{\boldsymbol{\theta}}^{\text{WLS}} - \frac{C\text{Tr}(\mathsf{B})}{\text{Tr}(\mathsf{B})^\top C \text{Tr}(\mathsf{B})}\left(\text{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\text{WLS}} - 1\right)$$

*where* $\widehat{\boldsymbol{\theta}}^{\text{WLS}}$ *is the WLS estimate* (12) *and* $C = (\mathbf{A}^\top \mathsf{W}\mathbf{A})^{-1}$. *The resulting MSE*

$$\mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\text{CWLS}}) = \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\text{CWLS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\text{CWLS}} - \boldsymbol{\theta})^\top = F \qquad (23)$$

*where* $F = C - \frac{C\text{Tr}(\mathsf{B})\text{Tr}(\mathsf{B})^\top C}{\text{Tr}(\mathsf{B})^\top C\text{Tr}(\mathsf{B})}$, *and* $\widehat{\boldsymbol{\theta}}^{\text{CWLS}}$ *is optimal in the sense that*

$$\mathbb{MSE}(\widehat{\boldsymbol{\theta}}) \ge \mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\text{CWLS}})$$

*where* $\widehat{\boldsymbol{\theta}}$ *is any unbiased estimate for* $\boldsymbol{\theta}$ *that is affine in* $\mathbf{y}$ *and satisfies the constraint* $\boldsymbol{\theta}^\top \text{Tr}(\mathsf{B}) = 1$.

### 3.4. Regularized weighted regression

Further, we introduce the following weighted regression with $\ell_2$-regularization:

$$\mathop{\text{minimize}}_{\boldsymbol{\theta}}\quad (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top \mathsf{W}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) + \gamma\|\boldsymbol{\theta}\|^2 \qquad (24a)$$

$$\text{subject to}\quad \boldsymbol{\theta}^\top \text{Tr}(\mathsf{B}) = 1. \qquad (24b)$$

where $\gamma \ge 0$ is a regularization parameter and $\|\cdot\|$ represents the 2-norm of a vector. The motivation for introducing (24) may arise from the following two aspects:

(i) When the POVM $\{\mathsf{M}_m\}_{m=1}^M$ is under-determinate, the matrix $\mathbf{A}$ in (4) might not have full column rank. As a result, the $\widehat{\boldsymbol{\theta}}^{\text{LS}}$, $\widehat{\boldsymbol{\theta}}^{\text{CLS}}$, and $\widehat{\boldsymbol{\theta}}^{\text{CWLS}}$ will all fail to produce a unique estimate to the quantum state. The additional $\ell_2$-regularization term in the prediction error function will resolve this non-uniqueness challenge.

(ii) In practice, the quantum state $\rho$ is often a combination of some finite number of pure states. As a result, a significant prior knowledge on $\rho$ would be that it is of low rank. Since the rank minimization optimization problem with convex constraints is NP-hard (Recht, Fazel, & Parrilo, 2010), the nuclear norm is a common alternative as an approximation of the rank constraint for matrices in various matrix optimization problems. Note that $\rho^\dagger\rho$ has the same rank as that of $\rho$. As a result, $\rho^\dagger\rho$ is still of low rank and its nuclear norm is

$$\begin{aligned}
\|\rho^\dagger\rho\|_\star &\triangleq \sum_{i=1}^{d}\sigma_i(\rho^\dagger\rho) = \text{Tr}(\rho^\dagger\rho) \\
&= \text{Tr}\left[\left(\sum_{i=1}^{d^2}\theta_i \mathsf{B}_i\right)^\dagger\left(\sum_{j=1}^{d^2}\theta_j \mathsf{B}_j\right)\right] \\
&= \sum_{i=1}^{d^2}|\theta_i|^2 \\
&= \|\boldsymbol{\theta}\|^2
\end{aligned}$$

where $\{\sigma_i(\rho^\dagger\rho), 1 \le i \le d\}$ are the singular values of $\rho^\dagger\rho$.

The two aspects are certainly connected in practice, where reconstruction of unknown low-rank quantum state is desired with a small number of measurement basis.

**Remark 1.** For the positive semidefinite quantum state $\rho$, penalizing the nuclear norm $\rho$ (see, e.g., Gross et al. (2010)) is not quite well-defined because

$$\begin{aligned}\|\rho\|_\star &\triangleq \sum_{i=1}^{d}\sigma_i(\rho) = \sum_{i=1}^{d}\sqrt{\lambda_i(\rho^\dagger\rho)}\\ &= \sum_{i=1}^{d}\lambda_i(\rho) = \mathrm{Tr}(\rho) = 1,\end{aligned} \tag{25}$$

where $\{\sigma_i(\rho), 1 \le i \le d\}$ and $\{\lambda_i(\rho), 1 \le i \le d\}$ are the singular values and eigenvalues of $\rho$, respectively. Note that (24) is essentially the regularized optimization approach adopted in Gross et al. (2010) for the numerical study of quantum state reconstruction problems.

**Remark 2.** The optimization problem (24) can be equivalently represented as

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad (\mathbf{y}-\mathbf{A}\boldsymbol{\theta})^\top \mathsf{W}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta}) \tag{26a}$$

$$\text{subject to} \quad \boldsymbol{\theta}^\top \mathrm{Tr}(\mathsf{B}) = 1, \ \|\boldsymbol{\theta}\|^2 \le c \tag{26b}$$

where $c > 0$ corresponds to $\gamma$. In (26b), it is clear that the $\ell_2$ norm of the $\boldsymbol{\theta}$ serves as a constraint from the two aspects of motivations for such regularization.

For convenience and consistence of the results displayed in the paper, here we first introduce the regularized weighted least squares (RWLS) estimate

$$\widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} \triangleq \arg\min_{\boldsymbol{\theta}}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta})^\top \mathsf{W}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta}) + \gamma\|\boldsymbol{\theta}\|^2 \tag{27a}$$

$$= (\mathbf{A}^\top \mathsf{W}\mathbf{A} + \gamma \mathsf{I})^{-1}\mathbf{A}^\top \mathsf{W}\mathbf{y}, \tag{27b}$$

where the constraint $\boldsymbol{\theta}^\top \mathrm{Tr}(\mathsf{B}) = 1$ is neglected.

The problem (24) also has a closed-from solution, which is stated in the following theorem.

**Theorem 2.** *The optimal weighted regularized quantum state estimate, denoted* $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$, *as the solution to (24) is given by*

$$\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - C\mathrm{Tr}(\mathsf{B})\frac{\mathrm{Tr}(\mathsf{B})^\top\widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - 1}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})} \tag{28}$$

*where* $C = (\mathbf{A}^\top \mathsf{W}\mathbf{A} + \gamma\mathsf{I})^{-1}$. *The resulting MSE matrix of* $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ *is*

$$\begin{aligned}\mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) &\triangleq \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})^\top\\ &= F - \gamma F(\mathsf{I} - \gamma\boldsymbol{\theta}\boldsymbol{\theta}^\top)F\end{aligned} \tag{29}$$

*where* $F = C - \frac{C\mathrm{Tr}(\mathsf{B})\mathrm{Tr}(\mathsf{B})^\top C}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}$.

It is worth noting that Theorem 2 does not depend on the rank of $\mathbf{A}$. The next theorem shows that the CRWLS estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ yields immediate improvement in terms of the MSE if the regularization parameter $\gamma$ is well chosen.

**Theorem 3.** *There holds*

$$\mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) < \mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\mathrm{CWLS}})$$

*if* $0 < \gamma < 2/\left(\|\boldsymbol{\theta}\|^2 - \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|^2}\right)$.

**Remark 3.** There holds from the Cauchy–Schwarz inequality that

$$\|\boldsymbol{\theta}\|^2\|\mathrm{Tr}(\mathsf{B})\|^2 \ge |\boldsymbol{\theta}^\top\mathrm{Tr}(\mathsf{B})|^2 = 1$$

for all quantum states $\rho$. Moreover, when strict equality takes place, there is $\lambda \in \mathbb{R}$ such that $\theta_i = \lambda\mathrm{Tr}(\mathsf{B}_i)$ for all $i = 1, \ldots, d^2$. As a result,

$$\mathrm{Tr}(\rho\mathsf{B}_i) = \theta_i = \lambda\mathrm{Tr}(\mathsf{B}_i), \ \ i = 1, \ldots, d^2$$

which implies $\rho = \lambda\mathsf{I}$, and hence $\lambda$ must be $1/d$. Therefore, we have just established that

$$\|\boldsymbol{\theta}\|^2 > \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|^2}$$

for all $\rho$ as quantum states except for $\rho = \mathsf{I}/d$.

Theorem 3 shows that regularization that considers the low rank property of the quantum state $\rho$ can further improve the estimate for the parameter vector $\boldsymbol{\theta}$ if we can choose a proper $\gamma$.

**Remark 4.** Theorem 1 indicates $\widehat{\boldsymbol{\theta}}^{\mathrm{CWLS}}$ has the smallest MSE among all the unbiased estimate of $\boldsymbol{\theta}$ affine with $\mathbf{y}$ while Theorem 3 shows that $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ has a smaller MSE than $\widehat{\boldsymbol{\theta}}^{\mathrm{CWLS}}$ even if $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ is also affine with $\mathbf{y}$. The reason is that regularization introduces a small bias but decreases the variance more such that the total MSE is smaller.

The estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ is a function of the regularization parameter $\gamma$, the selection of which needs to be determined carefully to achieve a desired performance. The essence of tuning $\gamma$ is to choose a proper model complexity for the estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ given the data. Here we provide a method of tuning $\gamma$ by the measurements based on the risk definition of the estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$. Also, we will prove that the tuning method is asymptotically optimal in the risk sense. For convenience of derivation, let us rewrite the estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ as the affine form with respective to the output $\mathbf{y}$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = \mathbf{H}\mathbf{y} + \boldsymbol{f} \tag{30}$$

with

$$\mathbf{H} = C\mathbf{A}^\top\mathsf{W} - C\mathrm{Tr}(\mathsf{B})\frac{\mathrm{Tr}(\mathsf{B})^\top C\mathbf{A}^\top\mathsf{W}}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}$$

$$\boldsymbol{f} = \frac{C\mathrm{Tr}(\mathsf{B})}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}$$

$$C = (\mathbf{A}^\top\mathsf{W}\mathbf{A} + \gamma\mathsf{I})^{-1}.$$

Let us introduce the risk for the estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ defined by Rao et al. (2008):

$$\begin{aligned}R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) &\triangleq \mathbb{E}\big(\mathbf{A}\boldsymbol{\theta} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}\big)^\top\mathsf{W}\big(\mathbf{A}\boldsymbol{\theta} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}\big)\\ &= \gamma^2\boldsymbol{\theta}^\top F\mathbf{A}^\top\mathsf{W}\mathbf{A}F\boldsymbol{\theta} + \mathrm{Tr}\left(F\mathbf{A}^\top\mathsf{W}\mathbf{A}F\mathbf{A}^\top\mathsf{W}\mathbf{A}\right)\end{aligned} \tag{31}$$

which is a reference measure to characterize how well the estimate (28) can be achieved, namely, gives a lower bound of the estimate (28) in the risk sense (31). Thus the regularization parameter $\gamma$ tuned by the risk (31)

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) \triangleq \arg\min_{\gamma\ge 0} R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) \tag{32}$$

is a theoretically optimal value for the regularization gain $\gamma$ for any given data in the risk sense. The cost function (31) of (32) requires the access to the true parameter $\boldsymbol{\theta}$, which is usually unavailable for a system to be identified.

In practice, we use an unbiased estimate for (31) as the cost function of an implementable tuning estimator in terms of data to estimate $\gamma$, which is given by

$$\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) \triangleq \arg\min_{\gamma\geq 0} (\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}})^\top \mathrm{W}(\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) + 2\mathrm{Tr}\big(\mathbf{AH}\big). \tag{33}$$

It can be verified that the expectation of the cost function (33) over the estimation error $\mathbf{e}$ is exactly the risk (31).

The properties of $\widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}})$ and $\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}})$ will be given in the next section under an alternative regression model.

## 4. An equivalent regression model

Up to now our discussions on the quantum state tomography problem have been around the linear model with an equality constraint:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \tag{34a}$$

$$\text{subject to } \boldsymbol{\theta}^\top \mathrm{Tr}(\mathsf{B}) = 1. \tag{34b}$$

In this section, we first present a way of transforming the constrained linear regression model into an unconstrained version. Then, under the new but equivalent model we establish some important asymptotic properties of the regularized regression solutions as the number of copies $n$ grows.

### 4.1. Eliminating equality constraint

Let us first construct an orthogonal matrix $\mathbf{Q}$ of size $d^2 \times d^2$ as follows: The first row of $\mathbf{Q}$ is $\mathrm{Tr}(\mathsf{B})^\top/\|\mathrm{Tr}(\mathsf{B})\|$ and the remaining rows are chosen such that $\mathbf{Q}$ is orthogonal. It follows from (34) that

$$\mathbf{y} = \mathbf{D}\boldsymbol{\beta} + \mathbf{e} \tag{35}$$

where

$$\mathbf{D} \triangleq \mathbf{A}\mathbf{Q}^\top = [\mathbf{d}, \mathbf{K}] \tag{36a}$$

$$\boldsymbol{\beta} \triangleq \mathbf{Q}\boldsymbol{\theta} = [\beta_1, \boldsymbol{\alpha}^\top]^\top. \tag{36b}$$

The constraint (34b) on $\boldsymbol{\theta}$ is forced by the fact that the first element $\beta_1$ of $\boldsymbol{\beta}$ is $1/\|\mathrm{Tr}(\mathsf{B})\|$. As a result, the problem (34) is equivalent to the unconstrained linear model

$$\mathbf{z} = \mathbf{y} - \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|}\mathbf{d} = \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}. \tag{37}$$

Clearly, by (36b)

$$\|\boldsymbol{\alpha}\|^2 = \|\boldsymbol{\theta}\|^2 - \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|^2}. \tag{38}$$

Thus, regularization (low rank property) on $\rho^\dagger\rho$ can also be embedded into $\boldsymbol{\alpha}$.

For the model (37), we can produce the corresponding LS, WLS, RWLS estimates. Here, we consider the RWLS estimate for (37) since other estimates (LS, WLS) are special cases by setting $\gamma = 0$ and/or $\mathrm{W} = \mathrm{I}$. The RWLS estimate for (37) is defined as

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} = \arg\min_{\boldsymbol{\alpha}} (\mathbf{z} - \mathbf{K}\boldsymbol{\alpha})^\top \mathrm{W}(\mathbf{z} - \mathbf{K}\boldsymbol{\alpha}) + \gamma\|\boldsymbol{\alpha}\|^2 \tag{39a}$$

$$= \mathbf{U}\mathbf{z} \tag{39b}$$

where

$$\mathbf{U} \triangleq \mathbf{V}\mathbf{K}^\top\mathrm{W}, \quad \mathbf{V} \triangleq (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}. \tag{40}$$

Intuitively, for an estimate $\widehat{\boldsymbol{\alpha}}$ of (37), the vector defined by

$$\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}) \triangleq \mathbf{Q}^\top \begin{bmatrix} \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} \tag{41}$$

should be the corresponding estimate for (34) and independent of the choice of $\mathbf{Q}$. However, this is not obvious. Now, we intend to show that the hypothesis above is true.

**Proposition 2.** *For any regularization parameter $\gamma \geq 0$, there holds*

$$\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) = \widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}. \tag{42}$$

*Moreover,*

$$\mathbb{MSE}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(\gamma)\big) \triangleq \mathbb{E}(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} - \boldsymbol{\alpha})(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} - \boldsymbol{\alpha})^\top = \gamma^2\mathbf{V}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{V} + \mathbf{V}\mathbf{K}^\top\mathrm{W}\mathbf{K}\mathbf{V}.$$

**Remark 5.** When $\gamma = 0$, the estimate $\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}$ is reduced to the WLS estimate of (37). Meanwhile, we have

$$\mathbb{MSE}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(\gamma)\big) < \mathbb{MSE}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(0)\big) \tag{43}$$

for $0 < \gamma < 2/\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$, an equivalent statement as Theorem 3.

### 4.2. Asymptotically optimal regularization gain

For the estimate (39), it also needs to well tune the regularization parameter $\gamma$. The risk for the estimate $\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}$ can be similarly defined as

$$R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) \triangleq \mathbb{E}\big(\mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}\big)^\top \mathrm{W}\big(\mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}\big) = \gamma^2\boldsymbol{\alpha}^\top\mathbf{V}\mathbf{K}^\top\mathrm{W}\mathbf{K}\mathbf{V}\boldsymbol{\alpha} + \mathrm{Tr}\big(\mathbf{V}\mathbf{K}^\top\mathrm{W}\mathbf{K}\mathbf{V}\mathbf{K}^\top\mathrm{W}\mathbf{K}\big) \tag{44}$$

and the resulting theoretically optimal regularization parameter is

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) \triangleq \arg\min_{\gamma\geq 0} R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(\gamma)). \tag{45}$$

Let us construct an unbiased estimate

$$\mathscr{C}_u(\gamma) \triangleq (\mathbf{z} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})^\top\mathrm{W}(\mathbf{z} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) + 2\mathrm{Tr}\big(\mathbf{K}\mathbf{U}\big) \tag{46}$$

for (44) and it can straightforwardly check its expectation with respect to $\mathbf{e}$ is $R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})$ up to a constant. Consequently, a practical regularization gain selection can be

$$\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) \triangleq \arg\min_{\gamma\geq 0} \mathscr{C}_u(\gamma) \tag{47}$$

which gives a way to estimate $\gamma$ directly by the data.

The following proposition illustrates that the tuning estimators (32) and (33) as well as (45) and (47) developed for the constrained model (34) and its unconstrained counterpart (37), respectively, are identical.

**Proposition 3.** *There hold*

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) = \widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) \tag{48a}$$

$$\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) = \widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}). \tag{48b}$$

Denote

$$\Sigma \triangleq \mathbf{K}^\top\mathrm{diag}\big([1/(p_1 - p_1^2), \ldots, 1/(p_M - p_M^2)]\big)\mathbf{K} \tag{49a}$$

$$\Upsilon \triangleq \mathbf{A}^\top\mathrm{diag}\big([1/(p_1 - p_1^2), \ldots, 1/(p_M - p_M^2)]\big)\mathbf{A}. \tag{49b}$$

We can establish the asymptotically optimal selection of the regularization parameter $\gamma$ explicitly for the regularized regression estimate of the quantum state in the risk senses (31) and (44).

**Theorem 4.** *Suppose* $\mathrm{rank}(\mathbf{A}) = d^2$. *As the number of copies $n \to \infty$, the limits take place*

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) \to \gamma^\star \text{ deterministically} \tag{50a}$$

$$\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\rm RWLS}) \to \gamma^\star \text{ almost surely} \tag{50b}$$

*where*

$$\gamma^\star = \frac{\mathrm{Tr}(\Sigma^{-1})}{\boldsymbol{\alpha}^\top \Sigma^{-1}\boldsymbol{\alpha}} = \frac{\mathrm{Tr}(\Upsilon^{-1}) - \frac{\mathrm{Tr(B)}^\top \Upsilon^{-2}\mathrm{Tr(B)}}{\mathrm{Tr(B)}^\top \Upsilon^{-1}\mathrm{Tr(B)}}}{\boldsymbol{\theta}^\top \Upsilon^{-1}\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^\top \Upsilon^{-1}\mathrm{Tr(B)}\mathrm{Tr(B)}^\top \Upsilon^{-1}\boldsymbol{\theta}}{\mathrm{Tr(B)}^\top \Upsilon^{-1}\mathrm{Tr(B)}}}$$

*is the asymptotically optimal selection of $\gamma$ for the estimate (28) of the quantum state in the risk senses (31) and (44). Moreover, there hold as $n \to \infty$*

$$n\left(\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\rm RWLS}) - \gamma^\star\right) \to \frac{3\gamma^\star\left(\gamma^\star\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-2}\boldsymbol{\alpha} - \mathrm{Tr}(\Sigma^{-2})\right)}{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} \tag{51}$$

*deterministically and*

$$\sqrt{n}\left(\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\rm RWLS}) - \gamma^\star\right) \to \mathscr{N}\left(0, \frac{(2\gamma^\star)^2\boldsymbol{\alpha}^\top \Sigma^{-3}\boldsymbol{\alpha}}{\left(\boldsymbol{\alpha}^\top \Sigma^{-1}\boldsymbol{\alpha}\right)^2}\right) \tag{52}$$

*in distribution.*

**Remark 6.** Theorem 4 shows that the implementable estimators $\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$ and $\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\rm RWLS})$ converge to the asymptotically optimal $\gamma^\star$ as the theoretically optimal estimators $\widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$ and $\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\rm RWLS})$ for any finite $n$ do. On the other hand, $\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$ and $\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\rm RWLS})$ have a slower rate of convergence than that of $\widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$ and $\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\rm RWLS})$.

## 5. Numerical examples

In Theorems 1–3, we have established the expressions and the resulting MSEs for $\widehat{\boldsymbol{\theta}}^{\rm CWLS}$ and $\widehat{\boldsymbol{\theta}}^{\rm CRWLS}$. Compared to the standard least squares solution $\widehat{\boldsymbol{\theta}}^{\rm LS}$ and weighted least squares solution $\widehat{\boldsymbol{\theta}}^{\rm WLS}$, $\widehat{\boldsymbol{\theta}}^{\rm CWLS}$ further makes use of the trace unity of the quantum state, while $\widehat{\boldsymbol{\theta}}^{\rm CRWLS}$ overcomes under-determinant measurement bases, provides reduction of MSE by tolerating estimation bias, and utilizes potential low-rank property of the quantum state. Theorem 4 further establishes that the practical computation of the regularization gain $\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$ is indeed asymptotically optimal. Meantime, the price for establishing $\widehat{\boldsymbol{\theta}}^{\rm CRWLS}$ (with asymptotically optimal $\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\rm CRWLS})$) is higher computation cost. In the following, we illustrate the connection and distinction of these estimators with a few concrete numerical examples.

All the numerical experiments are performed over Matlab 2014b for Mac OS X 10.14 with a 3.5 GHz Intel Core i5 CPU and an 8 GB RAM.

### 5.1. Overdeterminate measurement basis

**Example 1.** We consider the following quantum Werner state tomography for a two-qubit system (i.e., $d = 4$) as studied in Qi et al. (2013):

$$\rho_q = q|\Psi^-\rangle\langle\Psi^-| + \frac{1-q}{4}\mathrm{I}$$

where $|\Psi^-\rangle = (|01\rangle - |10\rangle)/\sqrt{2}$ and $q \in [0, 1]$ is a parameter characterizing different states. Particularly, $\rho_q$ changes from a completely mixed state ($q = 0$) to a pure state ($q = 1$) when $q$ varies from 0 to 1. Take an orthonormal basis $\{\mathrm{B}_i\}_{i=1}^{16}$ as

$$\mathrm{B}_i = \frac{1}{\sqrt{2}}\sigma_j \otimes \frac{1}{\sqrt{2}}\sigma_k, \quad i = 4j + k + 1$$

for $j, k = 0, 1, 2, 3$, where

$$\sigma_0 = I_2, \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

are the Pauli matrices. Let

$$|\varphi_1\rangle = \frac{1}{\sqrt{6}}[1, 1]^\top, |\varphi_2\rangle = \frac{1}{\sqrt{6}}[1, -1]^\top, |\varphi_3\rangle = \frac{1}{\sqrt{6}}[1, i]^\top,$$
$$|\varphi_4\rangle = \frac{1}{\sqrt{6}}[1, -i]^\top, |\varphi_5\rangle = \frac{1}{\sqrt{3}}[1, 0]^\top, |\varphi_6\rangle = \frac{1}{\sqrt{3}}[0, 1]^\top.$$

Then

$$\mathrm{E}_m = |\varphi_j\rangle\langle\varphi_j| \otimes |\varphi_k\rangle\langle\varphi_k|, \quad m = 6(j-1) + k,$$

for $j, k = 1, 2, \ldots, 6$ form our measurement basis $\{\mathrm{M}_m\}_{m=1}^{36}$ with $\mathrm{M}_m^\dagger = |\varphi_j\rangle \otimes |\varphi_k\rangle$. We can verify that the measurement set $\{\mathrm{M}_m\}_{m=1}^{36}$ is overcomplete and the matrix $\mathbf{A} = [\beta_1, \ldots, \beta_{36}]^\top \in \mathbb{R}^{36\times 16}$ has full column rank.

We first sample the parameter $q$ in the interval [0, 1] with the length 0.02 to generate 51 different quantum states $\rho_q$, and then for each $\rho_q$ perform the tomography procedure with $n = 110, 1100, 11\,000$ copies, respectively, to show the performance of the proposed estimators for the quantum states as the number of copies increases. The measurement process is simulated by i.i.d. multinormal random variable $\lambda$ taking values in the set $\{1, \ldots, 36\}$ with the probability $\mathbb{P}(\lambda = m) = p_m = \mathrm{Tr}(\mathrm{E}_m\rho_q)$, $1 \le m \le 36$ and by which the required copies for measurements are generated. Then the estimates $\widehat{p}_m$ are calculated by (2) from the measurement outcomes and hence the proposed estimators can be obtained by using $\widehat{p}_m$, $1 \le m \le 36$ and $\mathbf{A}$. Note that some of the $\widehat{p}_m$ might be or very close to zero, and whenever that happens we set $\widehat{p}_m = 10^{-8}$ in the weight matrix W for the sake of computation. In order to show the robustness of the proposed estimators against the errors, 1000 round experiments are executed.

Let us denote the standard, weighted, and constrained weighted estimates $\widehat{\boldsymbol{\theta}}^{\rm LS}(k)$, $\widehat{\boldsymbol{\theta}}^{\rm WLS}(k)$, $\widehat{\boldsymbol{\theta}}^{\rm CWLS}(k)$, and $\widehat{\boldsymbol{\theta}}^{\rm CRWLS}(k)$ (with $\gamma = 1/\left(\|\boldsymbol{\theta}\|^2 - \frac{1}{\|\mathrm{Tr(B)}\|^2}\right)$) according to (10), (12), (22), and (28) respectively, for the $k$th round, where W is replaced by its estimate (15). Note that the $\widehat{\boldsymbol{\theta}}^{\rm LS}(k)$ corresponds to the linear regression estimator proposed in Qi et al. (2013). In the following, we introduce two measures to evaluate the performance of the estimators. The first one is the experimental MSE $\mathbb{MSE}_{\rm exp}$ computed by averaging the squared errors from each round of experiments. The second one is the corresponding theoretical MSE given by (11), (13), (23), and (29), where the true W instead of its estimate is used. The two measures of the estimators for different quantum states $\rho_q$ with different copies are plotted in Figs. 1–3, respectively.

From these figures one can see that the experimental MSEs are approaching the theoretical MSEs and both of them decrease to zero as the number $n$ of copies grows large for all four estimates, LS, WLS, CWLS, and CRWLS. In particular, the CRWLS estimator has the highest tomography accuracy among all the estimators for large numbers of copies ($n = 1100, 11\,100$) as a confirmation of Theorem 3. For small $n = 110$, the WLS, CWLS, and CRWLS are apparently producing worse experimental MSE compared to the LS. The reason might be that the estimate $\widehat{\mathrm{W}}$ given by (15) constructed from the $\{\widehat{p}_m, 1 \le m \le M\}$ was not accurate enough for approximating the true W. For relatively larger $n = 11\,000$, the WLS, CWLS, and CRWLS all provide significant improvements compared to the LS. It is worth noting that even with small number of copies, the CRWLS may lead to drastically reduced error for small $q$, where $\rho_q$ tends to be closer to a completely mixed state.

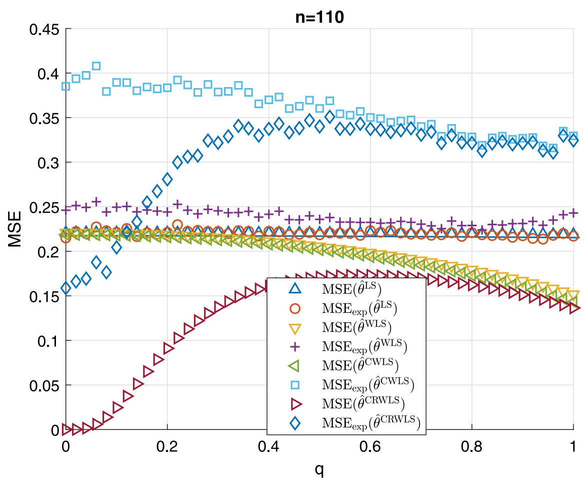



**Fig. 1.** MSEs for estimating Werner states with $n = 110$ copies.

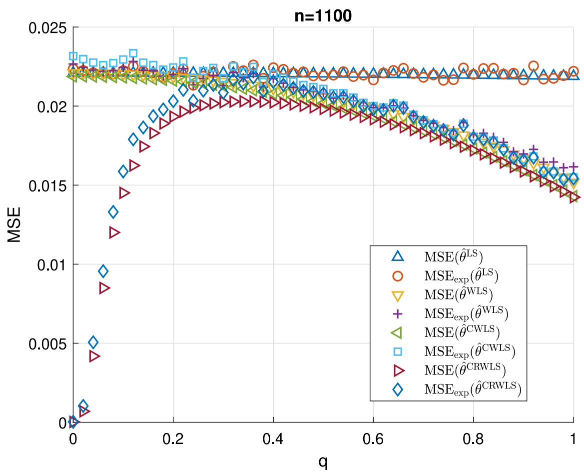



**Fig. 2.** MSEs for estimating Werner states with $n = 1100$ copies.

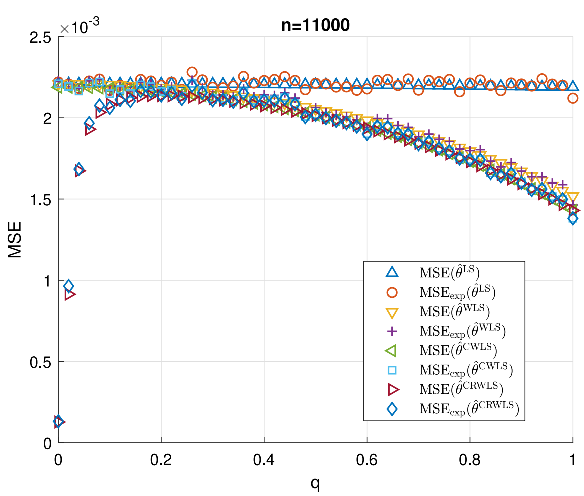



**Fig. 3.** MSEs for estimating Werner states with $n = 11\,000$ copies.

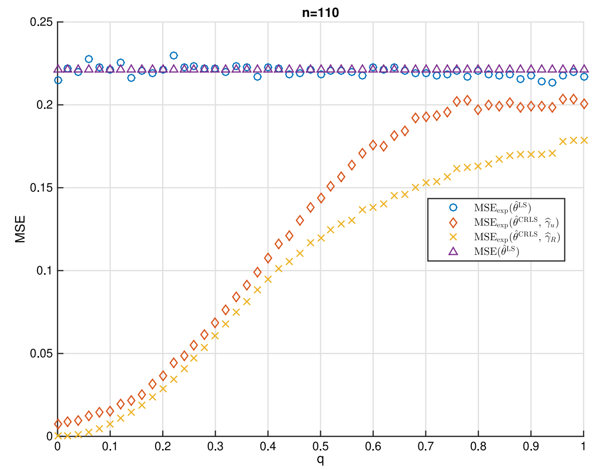



**Fig. 4.** CRLS vs. LS estimates for Werner states with $n = 110$ copies.

### 5.2. Small sample-size and optimal regularizer

As seen from Example 1, when the number of copies $n$ is small, the weighted estimates $\widehat{\boldsymbol{\theta}}^{\mathrm{WLS}}$, $\widehat{\boldsymbol{\theta}}^{\mathrm{CWLS}}$, and $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ involving the matrix $\widehat{\mathrm{W}}$ (15) may lead to lower accuracy compared to $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$. In the following example, we show that in this case forcing $\mathrm{W} = \mathrm{I}$ in $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}(k)$ to obtain a constrained regularized LS estimate (CRLS) would resolve the issue by developing the tuning methods similar to (32) and (33).

**Example 2.** Consider exactly the same quantum state and tomography setup as in Example 1. Let $\mathrm{W} = \mathrm{I}$ in $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ so that we define

$$\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}} = \widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}\Big|_{\mathrm{W}=\mathrm{I}}$$

as the CRLS estimate. For the estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}$, the regularization gain $\gamma$ is selected by the counterparts of (32) and (33) used for $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ as follows

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}) \triangleq \arg\min_{\gamma\geq 0} R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}) \tag{53}$$

$$\begin{aligned} R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}) &\triangleq \mathbb{E}\big(\mathbf{A}\boldsymbol{\theta} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}\big)^{\top}\big(\mathbf{A}\boldsymbol{\theta} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}\big) \\ &= \gamma^2\boldsymbol{\theta}^{\top}F\mathbf{A}^{\top}\mathbf{A}F\boldsymbol{\theta} + \mathrm{Tr}\left(\mathbf{A}F\mathbf{A}^{\top}\mathrm{P}\mathbf{A}F\mathbf{A}^{\top}\right) \end{aligned}$$

$$\widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}) \triangleq \arg\min_{\gamma\geq 0} \; \|\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}\|^2 + 2\mathrm{Tr}(\mathbf{AH}\mathrm{P}) \tag{54}$$

where $\mathbf{H}$ is exactly that given in (30) except for replacing W by I and P is defined in (11), under which for any $\rho_q$ we carry out the tomography procedure for 1000 rounds based on $n = 110, 1100$ copies, respectively. The resulting experimental MSEs $\mathbb{MSE}_{\mathrm{exp}}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}, \widehat{\gamma}_R(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}))$ and $\mathbb{MSE}_{\mathrm{exp}}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}, \widehat{\gamma}_u(\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}))$ are then computed as that in Example 1 and plotted in Figs. 4–5, respectively, in comparison to the experimental and theoretical MSEs of standard LS estimate $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$.

As seen from the numerical results, with $n = 110$, the regularizer for $\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}$ significantly improves the estimation accuracy compared to $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ under both $\widehat{\gamma}_R$ and $\widehat{\gamma}_u$. While with $n = 1100$, for relatively large $q$, the advantage of $\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}$ is no longer obvious compared to $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ while in this case, the use of the weight W becomes essential for further improving the performance. For illustrating the computational cost of calculating $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ and $\widehat{\boldsymbol{\theta}}^{\mathrm{CRLS}}$ with the $\gamma$ selection (54), the averaged running time for the cases $q = 0.1, 0.4, 0.6, 0.9$ and $n = 110, 1100$ is displayed in Table 1. Clearly, the improvement in estimation error from LS to CRLS is at the cost of lower computation efficiency.

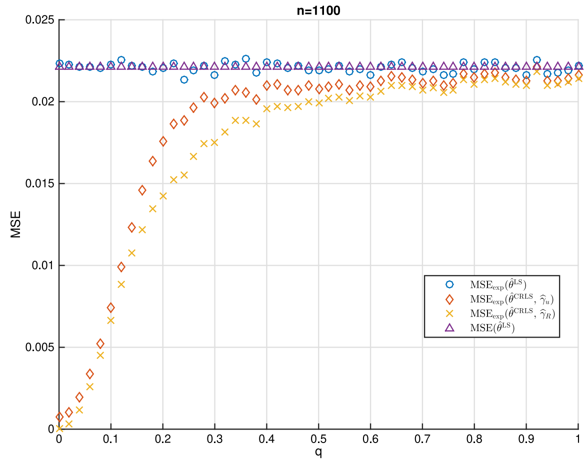



**Fig. 5.** CRLS vs. LS estimates for Werner states with $n = 1100$ copies.

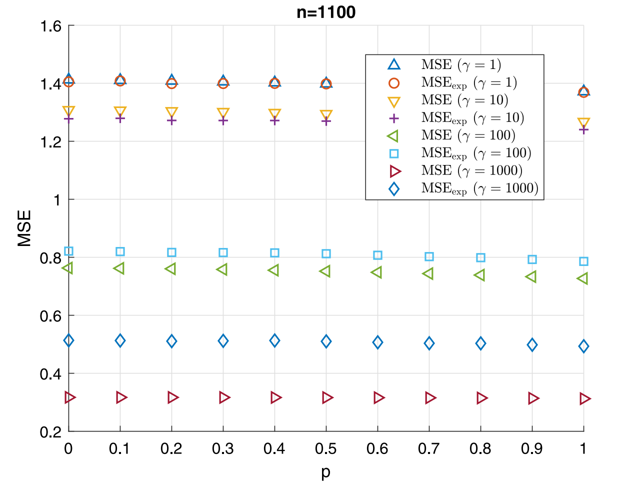



**Fig. 6.** MSEs for estimating the six-qubit state $\rho_p$ by CRWLS with $n = 1100$ copies.

**Table 1**
The averaged time for calculating $\widehat{\theta}^{\text{LS}}$ and $\widehat{\theta}^{\text{CRLS}}$.

| Time (s) | $q = 0.1$ | 0.4 | 0.6 | 0.9 |
|---|---|---|---|---|
| | | $n = 110$ | | |
| $\widehat{\theta}^{\text{LS}}$ | 3.97E−5 | 3.95E−5 | 3.94E−5 | 3.90E−5 |
| $\widehat{\theta}^{\text{CRLS}}$ with $\widehat{\gamma}_u$ | 0.0189 | 0.0162 | 0.0139 | 0.0121 |
| | | $n = 1100$ | | |
| $\widehat{\theta}^{\text{LS}}$ | 4.12E−5 | 4.10E−5 | 4.22E−5 | 4.14E−5 |
| $\widehat{\theta}^{\text{CRLS}}$ with $\widehat{\gamma}_u$ | 0.0172 | 0.0130 | 0.0141 | 0.0146 |

### *5.3. Under-determinate measurement basis*

**Example 3.** We consider a 6-qubit quantum state. We use the 6-qubit Pauli matrices to form our basis $\{\mathsf{B}_j, j = 1, \ldots, d^2\}$ with $d = 2^6 = 64$. Let $u = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2}i & \frac{\sqrt{3}}{2}i \end{bmatrix}$ be a $2 \times 2$ unitary matrix. Let

$$|\psi_1\rangle = [0, \ldots, 0, \overbrace{\sqrt{p}}^{42\text{-th}}, 0, \ldots, 0, \overbrace{\sqrt{1-p}}^{8\text{-th}}, 0, \ldots, 0]^\top,$$

$$|\psi_2\rangle = [0, \ldots, 0, \overbrace{1}^{59\text{-th}}, 0, \ldots, 0]^\top,$$

$$|\psi_3\rangle = [0, \ldots, 0, \overbrace{1}^{30\text{-th}}, 0, \ldots, 0]^\top,$$

be three pure states. Then $\rho_p = \left(u^{\otimes 6}\right)^\dagger \left(\frac{1}{3}|\psi_1\rangle\langle\psi_1| + \frac{1}{3}|\psi_2\rangle\langle\psi_2| + \frac{1}{3}|\psi_3\rangle\langle\psi_3|\right)u^{\otimes 6}$ is a rank-3 density matrix for a 6-qubit system for all $p \in [0, 1]$. Note that $\rho_p$ is low-rank but not sparse due to the existence of $u$. We index the set of 6-fold tensor product of Pauli matrices $\left\{\sigma_{l_1} \otimes \ldots \otimes \sigma_{l_6} : (l_1, \ldots, l_6) \in \{1, 2, 3\}^6\right\}$ by $\{P_j, j = 1, \ldots, 3^6\}$. The $P_j$'s are of full rank and have eigenvalues $\pm 1$. Denote by $Q_{j\pm}$ the projection onto the eigenspaces of $P_j$ with respect to $\pm 1$ respectively. We randomly choose and then fix $\{Q_{j_1+}, \ldots, Q_{j_{200}+}\}$ from $\{Q_{j+}, j = 1, \ldots, 3^6\}$. Then $\mathcal{M}^+ \triangleq \{M_k = \sqrt{Q_{j_k+}/200}, k = 1, \ldots, 200\}$ forms an under-complete measurement basis, and the resulting $\mathbf{A} = [\beta_1, \ldots, \beta_{200}]^\top \in \mathbb{R}^{200\times 4096}$ becomes under-determinate.

We use $n = 1100, 11\,000, 110\,000$ copies for each $\rho_p$ and perform independent measurements over each copy along any element in the basis $\mathcal{M}^+$. The parameter $p$ is sampled at $p = 0, 0.1, \ldots, 1$, and for each $\rho_p$, we carry out the tomography procedure and perform the $\widehat{\theta}^{\text{CRWLS}}$ estimator with $\gamma = 1, 10, 100, 1000$. Likewise, we test the $\widehat{\theta}^{\text{CRWLS}}$ estimator for 1000 independent rounds, whose experimental and theoretical MSEs $\mathbb{MSE}_{\text{exp}}$, $\mathbb{MSE}$ with respect to the parametrized states $\rho_p$ are plotted with $n = 1100, 11\,000, 110\,000$ in Figs. 6–8, respectively. From these plots we see that the MSE is fundamentally lower bounded by the $\mathcal{M}^+$ instead of heavily relying on the number of copies $n$. Moreover, the MSE and the risk are not quite sensitive with respect to the regularization gain $\gamma$. It is expected that these estimation results can be improved by utilizing the regularizer selection (33), but would require significantly higher computation cost. The computation time is approximately 0.6 s for all cases with different $p$, $\gamma$ and $n$ for the quantum state tomography procedure.

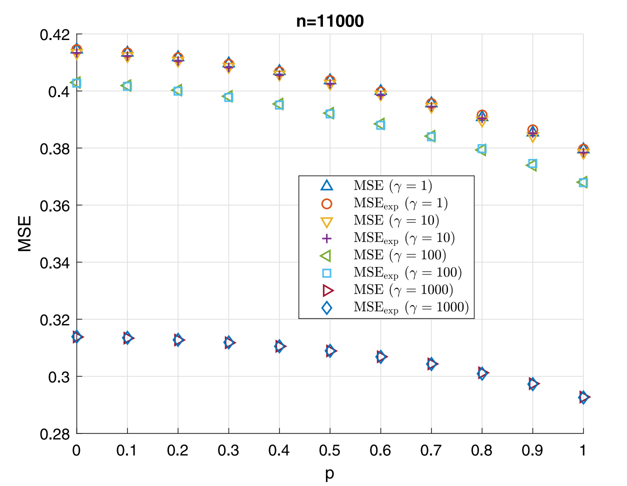



**Fig. 7.** MSEs for estimating the six-qubit state $\rho_p$ by CRWLS with $n = 11\,000$ copies.

## 6. Conclusions

We have studied a series of linear regression methods for quantum state tomography based on regularization. With complete or over-complete measurement bases, the empirical data

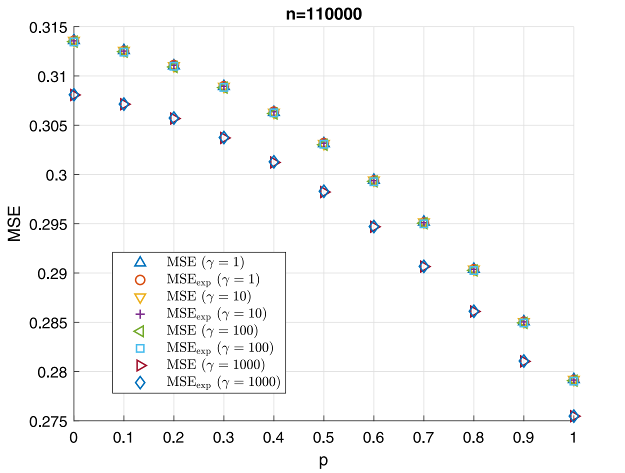



**Fig. 8.** MSEs for estimating the six-qubit state $\rho_p$ by CRWLS with $n = 110\,000$ copies.

was shown to be useful for the construction of a weighted LSE from the measurement outcomes of an unknown quantum state. It was proven that the trace-constrained weighted LSE is the optimal unbiased estimation among all linear estimators. For general measurement bases, either complete or incomplete, we showed that $\ell_2$-regularization with proper regularization parameter could yield even lower mean-square error under a penalty in bias. An explicit formula was established for the regularization parameter under an equivalent regression model, which shows that the proposed implementable tuning estimator is asymptotically optimal as the number of copies grows to infinity in the risk metric. An interesting future direction lies in regularization-based approach for Hamiltonian identification of quantum dynamical systems.

## Appendix A

The proofs of Proposition 1 and Theorem 1 are standard and can be found in the books (Amemiya, 1985; Theil, 1971). So, they are omitted for saving space.

### A.1. Proof of Theorem 2

The constrained optimization problem (24) is transformed to an unconstrained one by introducing Lagrange multiplier $\lambda$ and the resulting Lagrange function of (24) is

$$L(\boldsymbol{\theta}, \lambda) = (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^\top \mathsf{W}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) + \gamma\|\boldsymbol{\theta}\|^2 + \lambda(\boldsymbol{\theta}^\top \mathrm{Tr}(\mathsf{B}) - 1).$$

Thus the optimal solution $(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}, \lambda^*)$ of the problem (24) satisfies the first optimality condition

$$-2\mathbf{A}^\top \mathsf{W}(\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) + 2\gamma\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} + \lambda^* \mathrm{Tr}(\mathsf{B}) = 0 \tag{A.1a}$$

$$\mathrm{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = 1. \tag{A.1b}$$

It follows from (A.1a) that

$$\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - \lambda^* C\mathrm{Tr}(\mathsf{B})/2. \tag{A.2}$$

Furthermore, by using (A.1b), we have

$$\mathrm{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = \mathrm{Tr}(\mathsf{B})^\top (\widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - \lambda^* C\mathrm{Tr}(\mathsf{B})/2) = 1,$$

which yields

$$\lambda^* = \frac{2(\mathrm{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - 1)}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}. \tag{A.3}$$

Substituting (A.3) into (A.2) obtains

$$\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} = \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - C\mathrm{Tr}(\mathsf{B})\frac{\mathrm{Tr}(\mathsf{B})^\top \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - 1}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}.$$

Next, we compute the MSE matrix of $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$. By the constraint $\mathrm{Tr}(\mathsf{B})^\top \boldsymbol{\theta} = 1$, we have

$$\begin{aligned}
&\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta} \\
&= \widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - \boldsymbol{\theta} - \frac{C\mathrm{Tr}(\mathsf{B})}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}\big(\mathrm{Tr}(\mathsf{B})^\top(\widehat{\boldsymbol{\theta}}^{\mathrm{RWLS}} - \boldsymbol{\theta})\big) \\
&= C(-\gamma\boldsymbol{\theta} + \mathbf{A}^\top \mathsf{W}\mathbf{e}) - \frac{C\mathrm{Tr}(\mathsf{B})}{\mathrm{Tr}(\mathsf{B})^\top C\mathrm{Tr}(\mathsf{B})}\big(\mathrm{Tr}(\mathsf{B})^\top \\
&\quad C(-\gamma\boldsymbol{\theta} + \mathbf{A}^\top \mathsf{W}\mathbf{e})\big) \\
&= -\gamma F\boldsymbol{\theta} + F\mathbf{A}^\top \mathsf{W}\mathbf{e}.
\end{aligned}$$

The matrix $F$ has the following properties:

$$F^\top = F,\ F\mathrm{Tr}(\mathsf{B}) = 0,\ FC^{-1}F = F. \tag{A.4}$$

As a result, the MSE matrix of $\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}$ is

$$\begin{aligned}
&\mathbb{E}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})^\top \\
&= \gamma^2 F\boldsymbol{\theta}\boldsymbol{\theta}^\top F + F\mathbf{A}^\top \mathsf{W}\mathbf{A}F \\
&= \gamma^2 F\boldsymbol{\theta}\boldsymbol{\theta}^\top F + F(C^{-1} - \gamma \mathsf{I})F \\
&= FC^{-1}F - \gamma F(\mathsf{I} - \gamma\boldsymbol{\theta}\boldsymbol{\theta}^\top)F \\
&= F - \gamma F(\mathsf{I} - \gamma\theta\boldsymbol{\theta}^\top)F.
\end{aligned}$$

This completes the proof.

### A.2. Proof of Theorem 3

The proof requires the equivalent model (35) as well as resulting notations and conclusion in Section 4. Denote the orthogonal matrix appearing in (36) by

$$\mathbf{Q} = \begin{bmatrix} \frac{\mathrm{Tr}(\mathsf{B})^\top}{\|\mathrm{Tr}(\mathsf{B})\|} \\ \widetilde{\mathbf{Q}} \end{bmatrix}. \tag{A.5}$$

Thus by (36) and (42) we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta} &= \left[\frac{\mathrm{Tr}(\mathsf{B})}{\|\mathrm{Tr}(\mathsf{B})\|},\ \widetilde{\mathbf{Q}}^\top\right]\begin{bmatrix} \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|} \\ \widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} \end{bmatrix} \\
&\quad - \left[\frac{\mathrm{Tr}(\mathsf{B})}{\|\mathrm{Tr}(\mathsf{B})\|},\ \widetilde{\mathbf{Q}}^\top\right]\begin{bmatrix} \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|} \\ \boldsymbol{\alpha} \end{bmatrix} \\
&= \widetilde{\mathbf{Q}}^\top(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} - \boldsymbol{\alpha}).
\end{aligned} \tag{A.6}$$

When $0 < \gamma < 2/\big(\|\boldsymbol{\theta}\|^2 - \frac{1}{\|\mathrm{Tr}(\mathsf{B})\|^2}\big)$, there holds

$$\begin{aligned}
\mathbb{MSE}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}}) &\triangleq \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\mathrm{CRWLS}} - \boldsymbol{\theta})^\top \\
&= \widetilde{\mathbf{Q}}^\top \mathbb{E}\big((\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} - \boldsymbol{\alpha})(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}} - \boldsymbol{\alpha})\big)^\top \widetilde{\mathbf{Q}} \\
&= \widetilde{\mathbf{Q}}^\top \mathbb{MSE}(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(\gamma))\widetilde{\mathbf{Q}} \\
&< \widetilde{\mathbf{Q}}^\top \mathbb{MSE}(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(0))\widetilde{\mathbf{Q}} \\
&= \widetilde{\mathbf{Q}}^\top \mathbb{E}\big((\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(0) - \boldsymbol{\alpha})(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}(0) - \boldsymbol{\alpha})\big)^\top \widetilde{\mathbf{Q}}
\end{aligned}$$

$$= \mathbb{E}(\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}(0) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}(0) - \boldsymbol{\theta})^\top$$
$$= \mathbb{MSE}\left(\widehat{\boldsymbol{\theta}}^{\text{CWLS}}\right) \tag{A.7}$$

where the inequality is obtained by (43) in Section 4 and $\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}(0)$ is exactly $\widehat{\boldsymbol{\theta}}^{\text{CWLS}}$.

### A.3. Proof of Proposition 2 and Remark 5

It follows from Lemma B.1 that

$$\mathbf{Q}\widehat{\boldsymbol{\theta}}^{\text{CRWLS}} = \mathbf{QHy} + \mathbf{Qf}$$
$$= \begin{bmatrix} 0 \\ \mathbf{U} \end{bmatrix} \left( \mathbf{z} + \frac{\mathbf{d}}{\|\text{Tr}(\text{B})\|} \right) + \frac{1}{\|\text{Tr}(\text{B})\|} \begin{bmatrix} 1 \\ -\mathbf{Ud} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{\|\text{Tr}(\text{B})\|} \\ \widehat{\boldsymbol{\alpha}}^{\text{RWLS}} \end{bmatrix}. \tag{A.8}$$

Pre-multiplying $\mathbf{Q}^\top$ on both sides of (A.8) proves (42).

By (39), we have

$$\widehat{\boldsymbol{\alpha}}^{\text{RWLS}} - \boldsymbol{\alpha} = \mathbf{VK}^\top \text{W}\left(\mathbf{K}\boldsymbol{\alpha} + \mathbf{e}\right) - \boldsymbol{\alpha}$$
$$= \left(\mathbf{VK}^\top \text{WK} - \text{I}\right)\boldsymbol{\alpha} + \mathbf{VK}^\top \text{W}\mathbf{e}$$
$$= -\gamma \mathbf{V}\boldsymbol{\alpha} + \mathbf{VK}^\top \text{W}\mathbf{e}, \tag{A.9}$$

which derives

$$\mathbb{MSE}\left(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}(\gamma)\right) \triangleq \mathbb{E}(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}} - \boldsymbol{\alpha})(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}} - \boldsymbol{\alpha})^\top$$
$$= \gamma^2 \mathbf{V}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathbf{V} + \mathbf{VK}^\top \text{WKV}.$$

When $0 < \gamma < 2/\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$, we have $\gamma\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\text{I} < 0$ and further

$$\gamma\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \gamma(\mathbf{K}^\top \text{WK})^{-1} - 2\text{I} < 0$$

since $\gamma(\mathbf{K}^\top \text{WK})^{-1}$ is always positive definite. Thus we obtain

$$\mathbb{MSE}\left(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}(\gamma)\right) - \mathbb{MSE}\left(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}(0)\right)$$
$$= \gamma^2 \mathbf{V}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathbf{V} + \mathbf{VK}^\top \text{WKV} - \left(\mathbf{K}^\top \text{WK}\right)^{-1}$$
$$= \mathbf{V}\left(\gamma^2\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \mathbf{K}^\top \text{WK} - \mathbf{V}^{-1}\left(\mathbf{K}^\top \text{WK}\right)^{-1}\mathbf{V}^{-1}\right)\mathbf{V}$$
$$= \gamma \mathbf{V}\left(\gamma\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \gamma(\mathbf{K}^\top \text{WK})^{-1} - 2\text{I}\right)\mathbf{V}$$
$$< 0$$

if $0 < \gamma < 2/\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$.

### A.4. Proof of Proposition 3

By (A.8), we have

$$\mathbf{A}\boldsymbol{\theta} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\text{CRWLS}} = \mathbf{AQ}^\top \left(\mathbf{Q}\boldsymbol{\theta} - \mathbf{Q}\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}\right)$$
$$= \left[\mathbf{d}, \mathbf{K}\right] \left( \begin{bmatrix} \frac{1}{\|\text{Tr}(\text{B})\|} \\ \boldsymbol{\alpha} \end{bmatrix} - \begin{bmatrix} \frac{1}{\|\text{Tr}(\text{B})\|} \\ \widehat{\boldsymbol{\alpha}}^{\text{RWLS}} \end{bmatrix} \right)$$
$$= \mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}$$

which means that $R(\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}) = R(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}})$. So, the assertion (48a) holds.

Similarly, it gives

$$\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{\text{CRWLS}}$$
$$= \mathbf{z} + \frac{\mathbf{d}}{\|\text{Tr}(\text{B})\|} - \left[\mathbf{d}, \mathbf{K}\right] \begin{bmatrix} \frac{1}{\|\text{Tr}(\text{B})\|} \\ \widehat{\boldsymbol{\alpha}}^{\text{RWLS}} \end{bmatrix}$$
$$= \mathbf{z} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}. \tag{A.10}$$

The first equation in Lemma B.1 derives

$$\text{Tr}\left(\mathbf{AH}\right) = \text{Tr}\left(\mathbf{AQ}^\top \mathbf{QH}\right)$$
$$= \text{Tr}\left(\left[\mathbf{d}, \mathbf{K}\right] \begin{bmatrix} 0 \\ \mathbf{U} \end{bmatrix}\right) = \text{Tr}\left(\mathbf{KU}\right). \tag{A.11}$$

Combining (A.10) with (A.11) proves (48b).

### A.5. Proof of Theorem 4

We first prove the convergence (50). It follows from (A.9) that

$$\mathbb{E}\left(\mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}\right)^\top \text{W}\left(\mathbf{K}\boldsymbol{\alpha} - \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}\right)$$
$$= \gamma^2\boldsymbol{\alpha}^\top \mathbf{VK}^\top \text{WKV}\boldsymbol{\alpha} + \text{Tr}\left(\mathbf{VK}^\top \text{WKVK}^\top \text{WK}\right)$$

the two terms of which are

$$\gamma^2\boldsymbol{\alpha}^\top \mathbf{VK}^\top \text{WKV}\boldsymbol{\alpha}$$
$$= \gamma^2\boldsymbol{\alpha}^\top \left(\text{I} - \frac{\gamma}{n}\mathbf{S}\right)\mathbf{V}\boldsymbol{\alpha}$$
$$= \frac{\gamma^2}{n}\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} - \frac{\gamma^3}{n^2}\boldsymbol{\alpha}^\top \mathbf{S}^2\boldsymbol{\alpha} \tag{A.12a}$$

$$\mathbf{VK}^\top \text{WKVK}^\top \text{WK} = \left(\text{I} - \frac{\gamma}{n}\mathbf{S}\right)^2$$
$$= \text{I} - 2\frac{\gamma}{n}\mathbf{S} + \frac{\gamma^2}{n^2}\mathbf{S}^2 \tag{A.12b}$$

where $\mathbf{S}$ is defined in (B.3). Define

$$\mathscr{C}_R(\gamma) \triangleq n\left(R(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}) - \text{Tr}(\text{I})\right). \tag{A.13}$$

Thus

$$\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}) = \arg\min_{\gamma \geq 0} \mathscr{C}_R(\gamma). \tag{A.14}$$

Substituting (A.12) into (A.13) yields

$$\mathscr{C}_R(\gamma) = \gamma^2\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} - \frac{\gamma^3}{n}\boldsymbol{\alpha}^\top \mathbf{S}^2\boldsymbol{\alpha}$$
$$- 2\gamma\text{Tr}\left(\mathbf{S}\right) + \frac{\gamma^2}{n}\text{Tr}\left(\mathbf{S}^2\right)$$
$$\to \boldsymbol{\alpha}^\top \Sigma^{-1}\boldsymbol{\alpha}\gamma^2 - 2\text{Tr}\left(\Sigma^{-1}\right)\gamma \triangleq \mathscr{C}(\gamma) \tag{A.15}$$

as $n \to \infty$. It is clear that

$$\gamma^\star \triangleq \arg\min_{\gamma \geq 0} \mathscr{C}(\gamma) = \frac{\text{Tr}\left(\Sigma^{-1}\right)}{\boldsymbol{\alpha}^\top \Sigma^{-1}\boldsymbol{\alpha}}$$

which can also be expressed by $\boldsymbol{\theta}$ and $\Upsilon$ in terms of Lemma B.3.

By Lemma B.4, the limit $\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\text{RWLS}}) \to \gamma^\star$ holds as $n \to \infty$ since the convergence (A.15) is uniform over a compact subset of $[0, +\infty)$ that includes $\gamma^\star$.

For convenience of proving (52), denote

$$\widehat{\boldsymbol{\alpha}}^{\text{WLS}} \triangleq \mathbf{XK}^\top \text{W}\mathbf{z}, \quad \mathbf{X} \triangleq (\mathbf{K}^\top \text{WK})^{-1}. \tag{A.16}$$

It follows that

$$\mathbf{V} = \mathbf{X} - \gamma\mathbf{VX}, \quad \mathbf{VX}^{-1} = \text{I} - \gamma\mathbf{V} \tag{A.17}$$

$$\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{RWLS}} = \mathbf{KVK}^\top \text{W}\mathbf{z} = \mathbf{K}\left(\mathbf{X} - \gamma\mathbf{VX}\right)\mathbf{K}^\top \text{W}\mathbf{z}$$
$$= \mathbf{K}\widehat{\boldsymbol{\alpha}}^{\text{WLS}} - \gamma\mathbf{KV}\widehat{\boldsymbol{\alpha}}^{\text{WLS}}. \tag{A.18}$$

We first consider the decomposition of the first term of the cost function of (47)

$$
\begin{aligned}
&(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})^\top \mathrm{W}(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})\\
&=\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)^\top \mathrm{W}\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)\\
&\quad+\gamma^2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}\mathbf{K}^\top \mathrm{W}\mathbf{K}\mathbf{V}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad+2\gamma(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}\mathbf{K}^\top \mathrm{W}\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big).
\end{aligned}
\tag{A.19}
$$

The second term of (A.19) is

$$
\begin{aligned}
&\gamma^2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}\mathbf{K}^\top \mathrm{W}\mathbf{K}\mathbf{V}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&=\gamma^2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\big(\mathrm{I}-\gamma\mathbf{V}\big)\mathbf{V}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&=\gamma^2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})-\gamma^3(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}^2\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&=\frac{\gamma^2}{n}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})-\frac{\gamma^3}{n^2}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}^2\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}.
\end{aligned}
\tag{A.20}
$$

The third term of (A.19) is

$$
\begin{aligned}
&2\gamma(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}\mathbf{K}^\top \mathrm{W}\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)\\
&\quad=2\gamma(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{V}\big(\mathbf{K}^\top \mathrm{W}\mathbf{z}-\mathbf{K}^\top \mathrm{W}\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)\\
&\quad=0.
\end{aligned}
\tag{A.21}
$$

Further, we have

$$
\mathbf{K}\mathbf{U}=\mathbf{V}\mathbf{K}^\top \mathrm{W}\mathbf{K}=\mathrm{I}-\frac{\gamma}{n}\mathbf{S}.
\tag{A.22}
$$

Define

$$
\begin{aligned}
\mathscr{C}_U(\gamma)\triangleq n\Big(&\mathscr{C}_u(\gamma)-\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)^\top \mathrm{W}\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)\\
&-2\mathrm{Tr}(\mathrm{I})\Big).
\end{aligned}
$$

Thus

$$
\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})=\arg\min_{\gamma\ge 0}\mathscr{C}_U(\gamma)
\tag{A.23}
$$

since $\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)^\top \mathrm{W}\big(\mathbf{z}-\mathbf{K}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\big)$ is independent of $\gamma$. Substituting (A.20)–(A.22) into (A.23) turns out

$$
\begin{aligned}
\mathscr{C}_U(\gamma)&=\gamma^2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\frac{\gamma^3}{n}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}^2\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad-2\gamma\mathrm{Tr}\,(\mathbf{S})\\
&\quad\to\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}\gamma^2-2\mathrm{Tr}\big(\Sigma^{-1}\big)\gamma=\mathscr{C}(\gamma)
\end{aligned}
\tag{A.24}
$$

as $n\to\infty$ since $\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\to\boldsymbol{\alpha}$ almost surely as $n\to\infty$. It follows from Lemma B.4 that the limit (52) is true since the convergence (A.24) is uniform over a compact subset of $[0,+\infty)$ that includes $\gamma^\star$.

It remains to show the rates of convergence (51) and (52).

For this, we first calculate the first and second order derivatives of $\mathscr{C}_R(\gamma)$ with respective to $\gamma$ with the help of Lemma B.2:

$$
\begin{aligned}
\frac{\mathrm{d}\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma}&=2\gamma\boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha}-\frac{\gamma^2}{n}\boldsymbol{\alpha}^\top\mathbf{S}^2\boldsymbol{\alpha}\\
&\quad-\frac{3\gamma^2}{n}\boldsymbol{\alpha}^\top\mathbf{S}^2\boldsymbol{\alpha}+\frac{2\gamma^3}{n^2}\boldsymbol{\alpha}^\top\mathbf{S}^3\boldsymbol{\alpha}\\
&\quad-2\mathrm{Tr}\big(\mathbf{S}\big)+\frac{2\gamma}{n}\mathrm{Tr}\big(\mathbf{S}^2\big)\\
&\quad+\frac{2\gamma}{n}\mathrm{Tr}\big(\mathbf{S}^2\big)-\frac{2\gamma^2}{n^2}\mathrm{Tr}\big(\mathbf{S}^3\big)\\
\frac{\mathrm{d}^2\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma^2}&=2\boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha}+O\left(\frac{1}{n}\right).
\end{aligned}
$$

By using a Taylor expansion, we have

$$
\begin{aligned}
0=\frac{\mathrm{d}\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})}&=\frac{\mathrm{d}\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\gamma^\star}\\
&\quad+\frac{\mathrm{d}^2\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma^2}\Big|_{\gamma=\overline{\gamma}}\big(\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})-\gamma^\star\big).
\end{aligned}
$$

where $\overline{\gamma}$ is a real number between $\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})$ and $\gamma^\star$, which implies that

$$
\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})-\gamma^\star=-\left(\frac{\mathrm{d}^2\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma^2}\Big|_{\gamma=\overline{\gamma}}\right)^{-1}\frac{\mathrm{d}\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\gamma^\star}.
$$

As $n\to\infty$, we have

$$
\begin{aligned}
n\frac{\mathrm{d}\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\gamma^\star}&=2n\big(\gamma^\star\boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha}-\mathrm{Tr}\big(\mathbf{S}\big)\big)\\
&\quad-4(\gamma^\star)^2\boldsymbol{\alpha}^\top\mathbf{S}^2\boldsymbol{\alpha}+4\gamma^\star\mathrm{Tr}\big(\mathbf{S}^2\big)+O\left(\frac{1}{n}\right)\\
&=-2(\gamma^\star)^2\boldsymbol{\alpha}^\top\mathbf{S}\Sigma^{-1}\boldsymbol{\alpha}+2\gamma^\star\mathrm{Tr}\big(\mathbf{S}\Sigma^{-1}\big)\\
&\quad-4(\gamma^\star)^2\boldsymbol{\alpha}^\top\mathbf{S}^2\boldsymbol{\alpha}+4\gamma^\star\mathrm{Tr}\big(\mathbf{S}^2\big)+O\left(\frac{1}{n}\right)\\
&\to-6(\gamma^\star)^2\boldsymbol{\alpha}^\top\Sigma^{-2}\boldsymbol{\alpha}+6\gamma^\star\mathrm{Tr}\big(\Sigma^{-2}\big)\\
\frac{\mathrm{d}^2\,\mathscr{C}_R(\gamma)}{\mathrm{d}\,\gamma^2}\Big|_{\gamma=\overline{\gamma}}&\to 2\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}
\end{aligned}
$$

which yields

$$
\begin{aligned}
&n\big(\widehat{\gamma}_R(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})-\gamma^\star\big)\\
&\quad\to\frac{3\gamma^\star\big(\gamma^\star\boldsymbol{\alpha}^\top\Sigma^{-2}\boldsymbol{\alpha}-\mathrm{Tr}\big(\Sigma^{-2}\big)\big)}{\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}}.
\end{aligned}
\tag{A.27}
$$

For proving (52), the procedure is similar. By Lemma B.2, the first and second derivatives of $\mathscr{C}_U(\gamma)$ are

$$
\begin{aligned}
\frac{\mathrm{d}\,\mathscr{C}_U(\gamma)}{\mathrm{d}\,\gamma}&=2\gamma(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\frac{\gamma^2}{n}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}^2\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad-\frac{3\gamma^2}{n}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}^2\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}+\frac{2\gamma^3}{n^2}(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}^3\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad-2\mathrm{Tr}\big(\mathbf{S}\big)+\frac{2\gamma}{n}\mathrm{Tr}\big(\mathbf{S}^2\big)\\
\frac{\mathrm{d}^2\,\mathscr{C}_U(\gamma)}{\mathrm{d}\,\gamma^2}&=2(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}+O_p\left(\frac{1}{n}\right).
\end{aligned}
$$

Applying the Taylor expansion obtains

$$
\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})-\gamma^\star=-\left(\frac{\mathrm{d}^2\,\mathscr{C}_U(\gamma)}{\mathrm{d}\,\gamma^2}\Big|_{\gamma=\overline{\gamma}}\right)^{-1}\frac{\mathrm{d}\,\mathscr{C}_U(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\gamma^\star}
$$

where $\overline{\gamma}$ is a real number between $\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}})$ and $\gamma^\star$. By a straightforward calculation, we have

$$
\sqrt{n}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}\big)\to\mathscr{N}\left(0,\Sigma^{-1}\right)
$$

by further using the Delta method,

$$
\begin{aligned}
&\sqrt{n}\big((\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top\mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}\big)\\
&=\sqrt{n}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}\big)^\top\mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad+\sqrt{n}\boldsymbol{\alpha}^\top\big(\mathbf{S}-\Sigma^{-1}\big)\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad+\boldsymbol{\alpha}^\top\Sigma^{-1}\sqrt{n}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}\big)\\
&=\sqrt{n}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}\big)^\top\mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}\\
&\quad+\boldsymbol{\alpha}^\top\Sigma^{-1}\sqrt{n}\big(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}}-\boldsymbol{\alpha}\big)+O_p\left(\frac{1}{\sqrt{n}}\right)\\
&\to\mathscr{N}\left(0,4\,\boldsymbol{\alpha}^\top\Sigma^{-3}\boldsymbol{\alpha}\right).
\end{aligned}
$$

It follows that

$$\sqrt{n}\frac{\mathrm{d}\,\mathscr{C}_U(\gamma)}{\mathrm{d}\,\gamma}\Big|_{\gamma=\gamma^\star}$$
$$= 2\gamma^\star(\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}} - 2\mathrm{Tr}(\mathbf{S}) + O_p\left(\frac{1}{\sqrt{n}}\right)$$
$$= 2\sqrt{n}\gamma^\star\left((\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}} - \boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}\right)$$
$$+ 2\sqrt{n}\big(\mathrm{Tr}(\Sigma^{-1}) - 2\mathrm{Tr}(\mathbf{S})\big) + O_p\left(\frac{1}{\sqrt{n}}\right)$$
$$= 2\gamma^\star\sqrt{n}\left((\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}})^\top \mathbf{S}\widehat{\boldsymbol{\alpha}}^{\mathrm{WLS}} - \boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha}\right) + O_p\left(\frac{1}{\sqrt{n}}\right)$$
$$\to \mathscr{N}\left(0, 16(\gamma^\star)^2\boldsymbol{\alpha}^\top\Sigma^{-3}\boldsymbol{\alpha}\right).$$

As a result,

$$\sqrt{n}\big(\widehat{\gamma}_u(\widehat{\boldsymbol{\alpha}}^{\mathrm{RWLS}}) - \gamma^\star\big) \to \mathscr{N}\left(0, \frac{(2\gamma^\star)^2\boldsymbol{\alpha}^\top\Sigma^{-3}\boldsymbol{\alpha}}{(\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha})^2}\right).$$

## Appendix B

This appendix contains the technical lemmas used in the proof in Appendix A.

**Lemma B.1.** *We have*

$$\mathbf{QH} = \begin{bmatrix} 0 \\ \mathbf{U} \end{bmatrix}, \quad \mathbf{Qf} = \frac{1}{\|\mathrm{Tr}(\mathrm{B})\|}\begin{bmatrix} 1 \\ -\mathbf{Ud} \end{bmatrix}. \tag{B.1}$$

**Proof.** For convenience of proof, we denote

$$C = (\mathbf{A}^\top \mathrm{W}\mathbf{A} + \gamma\mathrm{I})^{-1}$$
$$\mathbf{G} = (\mathbf{D}^\top \mathrm{W}\mathbf{D} + \gamma\mathrm{I})^{-1}$$

as well as the column vector $\boldsymbol{e}_1$ of length $d^2$ and matrix $\boldsymbol{e}_{-1}$ of size $d^2$ by $d^2 - 1$ that make up the identity matrix

$$\begin{bmatrix} \boldsymbol{e}_1, & \boldsymbol{e}_{-1} \end{bmatrix} = \mathrm{I}.$$

Thus the following identities hold.

$$\mathbf{G} = \mathbf{Q}C\mathbf{Q}^\top, \quad \mathbf{D} = \mathbf{A}\mathbf{Q}^\top = [\mathbf{d}, \mathbf{K}] \tag{B.2a}$$
$$\boldsymbol{e}_1^\top \mathbf{Q} = \mathrm{Tr}(\mathrm{B})^\top/\|\mathrm{Tr}(\mathrm{B})\|, \quad \mathbf{Q}\mathrm{Tr}(\mathrm{B}) = \|\mathrm{Tr}(\mathrm{B})\|\boldsymbol{e}_1 \tag{B.2b}$$
$$\boldsymbol{e}_{-1}^\top \mathbf{G}\boldsymbol{e}_1/\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1 = -(\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\mathbf{K}^\top\mathrm{W}\mathbf{d}. \tag{B.2c}$$

The identities (B.2a)–(B.2b) can be verified by a straightforward calculation.

Eq. (B.2c) is obtained by choosing the (2, 1)-block submatrix of the identity

$$\mathbf{G}^{-1}\mathbf{G} = \begin{bmatrix} \mathbf{d}^\top\mathrm{W}\mathbf{d} + \gamma & \mathbf{d}^\top\mathrm{W}\mathbf{K} \\ \mathbf{K}^\top\mathrm{W}\mathbf{d} & \mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I} \end{bmatrix} \times \begin{bmatrix} \boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1 & \boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_{-1} \\ \boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_1 & \boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_{-1} \end{bmatrix} = \mathrm{I}.$$

By using (B.2a)–(B.2b), we have

$$\mathbf{QH} = \mathbf{Q}C\mathbf{A}^\top\mathrm{W} - \mathbf{Q}C\mathrm{Tr}(\mathrm{B})\frac{\mathrm{Tr}(\mathrm{B})^\top C\mathbf{A}^\top\mathrm{W}}{\mathrm{Tr}(\mathrm{B})^\top C\mathrm{Tr}(\mathrm{B})}$$
$$= \mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathbf{A}^\top\mathrm{W}$$
$$- \mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathrm{Tr}(\mathrm{B})\frac{\mathrm{Tr}(\mathrm{B})^\top\mathbf{Q}^\top\mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathbf{A}^\top\mathrm{W}}{\mathrm{Tr}(\mathrm{B})^\top\mathbf{Q}^\top\mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathrm{Tr}(\mathrm{B})}$$
$$= \mathbf{G}\mathbf{D}\mathrm{W} - \mathbf{G}\boldsymbol{e}_1\boldsymbol{e}_1^\top\mathbf{G}\mathbf{D}\mathrm{W}/\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1.$$

It is clear that

$$\boldsymbol{e}_1^\top\mathbf{QH} = \boldsymbol{e}_1^\top\mathbf{G}\mathbf{D}\mathrm{W} - \boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1\boldsymbol{e}_1^\top\mathbf{G}\mathbf{D}\mathrm{W}/\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1 = 0.$$

and further it follows from (B.2c) that

$$\boldsymbol{e}_{-1}^\top\mathbf{QH} = \boldsymbol{e}_{-1}^\top\mathbf{G}\mathbf{D}\mathrm{W} - \boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_1\boldsymbol{e}_1^\top\mathbf{G}\mathbf{D}\mathrm{W}/\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1$$
$$= \left(\boldsymbol{e}_{-1}^\top - \frac{\boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1}\boldsymbol{e}_1^\top\right)\mathbf{G}\mathbf{D}\mathrm{W}$$
$$= \begin{bmatrix} -\frac{\boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1}, & \mathrm{I} \end{bmatrix}\begin{bmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{e}_{-1}^\top \end{bmatrix}\mathbf{G}\mathbf{D}\mathrm{W}$$
$$= \begin{bmatrix} (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\mathbf{K}^\top\mathrm{W}\mathbf{d}, & \mathrm{I} \end{bmatrix}\mathbf{G}\mathbf{D}^\top\mathrm{W}$$
$$= (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\begin{bmatrix} \mathbf{K}^\top\mathrm{W}\mathbf{d}, & \mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I} \end{bmatrix}\mathbf{G}\mathbf{D}^\top\mathrm{W}$$
$$= (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\boldsymbol{e}_{-1}^\top\mathbf{G}^{-1}\mathbf{G}\mathbf{D}^\top\mathrm{W}$$
$$= (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\boldsymbol{e}_{-1}^\top\mathbf{D}^\top\mathrm{W}$$
$$= (\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\mathbf{K}^\top\mathrm{W}$$
$$= \mathbf{U}.$$

Using (B.2a)–(B.2b) again, one yields

$$\mathbf{Qf} = \frac{\mathbf{Q}C\mathrm{Tr}(\mathrm{B})}{\mathrm{Tr}(\mathrm{B})^\top C\mathrm{Tr}(\mathrm{B})}$$
$$= \frac{\mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathrm{Tr}(\mathrm{B})}{\mathrm{Tr}(\mathrm{B})^\top\mathbf{Q}^\top\mathbf{Q}C\mathbf{Q}^\top\mathbf{Q}\mathrm{Tr}(\mathrm{B})}$$
$$= \frac{\mathbf{G}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1\|\mathrm{Tr}(\mathrm{B})\|}.$$

Thus by (B.2c) we have

$$\boldsymbol{e}_1^\top\mathbf{Qf} = \frac{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1\|\mathrm{Tr}(\mathrm{B})\|} = 1/\|\mathrm{Tr}(\mathrm{B})\|$$
$$\boldsymbol{e}_{-1}^\top\mathbf{Qf} = \frac{\boldsymbol{e}_{-1}^\top\mathbf{G}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top\mathbf{G}\boldsymbol{e}_1\|\mathrm{Tr}(\mathrm{B})\|}$$
$$= -(\mathbf{K}^\top\mathrm{W}\mathbf{K} + \gamma\mathrm{I})^{-1}\mathbf{K}^\top\mathrm{W}\mathbf{d}/\|\mathrm{Tr}(\mathrm{B})\|$$
$$= -\mathbf{Ud}/\|\mathrm{Tr}(\mathrm{B})\|.$$

This completes the proof. ■

**Lemma B.2.** *Denote*

$$\mathbf{S} = \left(\Sigma + \frac{\gamma}{n}\mathrm{I}\right)^{-1}. \tag{B.3}$$

*We have*

$$\frac{\mathrm{d}\,\boldsymbol{g}^\top\mathbf{S}\boldsymbol{g}}{\mathrm{d}\,\gamma} = -\frac{\boldsymbol{g}^\top\mathbf{S}^2\boldsymbol{g}}{n}, \quad \frac{\mathrm{d}\,\boldsymbol{g}^\top\mathbf{S}^2\boldsymbol{g}}{\mathrm{d}\,\gamma} = -\frac{2\boldsymbol{g}^\top\mathbf{S}^3\boldsymbol{g}}{n} \tag{B.4}$$
$$\frac{\mathrm{d}\mathrm{Tr}(\mathbf{S})}{\mathrm{d}\,\gamma} = -\frac{\mathrm{Tr}(\mathbf{S}^2)}{n}, \quad \frac{\mathrm{d}\mathrm{Tr}(\mathbf{S}^2)}{\mathrm{d}\,\gamma} = -\frac{2\mathrm{Tr}(\mathbf{S}^3)}{n}. \tag{B.5}$$

*where* $\boldsymbol{g}$ *is any column vector.*

The proof is carried out by making use of the matrix differentiation formulas in Petersen and Pedersen (2012, Chapter 2) and is omitted due to limited space.

**Lemma B.3.** *We have*

$$\mathrm{Tr}(\Sigma^{-1}) = \Upsilon^{-1} - \frac{\mathrm{Tr}(\mathrm{B})^\top\Upsilon^{-2}\mathrm{Tr}(\mathrm{B})}{\mathrm{Tr}(\mathrm{B})^\top\Upsilon^{-1}\mathrm{Tr}(\mathrm{B})}$$
$$\boldsymbol{\alpha}^\top\Sigma^{-1}\boldsymbol{\alpha} = \boldsymbol{\theta}^\top\Upsilon^{-1}\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^\top\Upsilon^{-1}\mathrm{Tr}(\mathrm{B})\mathrm{Tr}(\mathrm{B})^\top\Upsilon^{-1}\boldsymbol{\theta}}{\mathrm{Tr}(\mathrm{B})^\top\Upsilon^{-1}\mathrm{Tr}(\mathrm{B})}.$$

**Proof.** By letting $\gamma = 0$ in Lemma B.1, we have

$$\mathbf{QH} = \begin{bmatrix} 0 \\ (\mathbf{K}^\top\mathrm{W}\mathbf{K})^{-1}\mathbf{K}^\top\mathrm{W} \end{bmatrix} \tag{B.6}$$

where

$$\mathbf{H} = C\mathbf{A}^\top \mathrm{W} - C\mathrm{Tr(B)}\frac{\mathrm{Tr(B)}^\top C\mathbf{A}^\top \mathrm{W}}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}}$$

$$C = \left(\mathbf{A}^\top \mathrm{W}\mathbf{A}\right)^{-1}.$$

It follows that

$$\begin{aligned}
&\mathbf{Q}^\top \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{K}^\top \mathrm{W}\mathbf{K})^{-1} \end{bmatrix} \mathbf{Q} = \mathbf{H}\mathrm{W}^{-1}\mathbf{H}^\top \\
&= \left(\mathrm{I} - \frac{C\mathrm{Tr(B)}\mathrm{Tr(B)}^\top}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}}\right) C \left(\mathrm{I} - \frac{\mathrm{Tr(B)}\mathrm{Tr(B)}^\top C}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}}\right) \\
&= \left(C - \frac{C\mathrm{Tr(B)}\mathrm{Tr(B)}^\top C}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}}\right)
\end{aligned}$$

which derives that

$$\begin{aligned}
\mathrm{Tr}\left(\Sigma^{-1}\right) &= n\mathrm{Tr}\left((\mathbf{K}^\top \mathrm{W}\mathbf{K})^{-1}\right) \\
&= n\mathrm{Tr}\left(C\right) - \frac{n\mathrm{Tr(B)}^\top C^2 \mathrm{Tr(B)}}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}} \\
&= \mathrm{Tr}\left(\Upsilon^{-1}\right) - \frac{\mathrm{Tr(B)}^\top \Upsilon^{-2}\mathrm{Tr(B)}}{\mathrm{Tr(B)}^\top \Upsilon^{-1}\mathrm{Tr(B)}} \\
\boldsymbol{\alpha}^\top \Sigma^{-1}\boldsymbol{\alpha} &= n\boldsymbol{\alpha}^\top (\mathbf{K}^\top \mathrm{W}\mathbf{K})^{-1}\boldsymbol{\alpha} \\
&= n\boldsymbol{\theta}^\top \mathbf{Q}^\top \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{K}^\top \mathrm{W}\mathbf{K})^{-1} \end{bmatrix} \mathbf{Q}\boldsymbol{\theta} \\
&= n\boldsymbol{\theta}^\top \left(C - \frac{C\mathrm{Tr(B)}\mathrm{Tr(B)}^\top C}{\mathrm{Tr(B)}^\top C\mathrm{Tr(B)}}\right)\boldsymbol{\theta} \\
&= \boldsymbol{\theta}^\top \Upsilon^{-1}\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^\top \Upsilon^{-1}\mathrm{Tr(B)}\mathrm{Tr(B)}^\top \Upsilon^{-1}\boldsymbol{\theta}}{\mathrm{Tr(B)}^\top \Upsilon^{-1}\mathrm{Tr(B)}}
\end{aligned}$$

This completes the proof. ■

**Lemma B.4** (*Ljung, 1999, Theorem 8.2*). *Assume that*

*(1) $\mathscr{C}(\gamma)$ is a deterministic function that is continuous in $\gamma \in \Omega$ and minimized at the point $\gamma^\star$, where $\Omega$ is a compact subset of $\mathbb{R}$.*

*(2) A sequence of functions $\{\mathscr{C}_n(\gamma)\}$ converges to $\mathscr{C}(\gamma)$ almost surely and uniformly in $\Omega$ as $n$ goes to $\infty$.*

*Then $\widehat{\gamma}_n = \arg\min_{\gamma\in\Omega} \mathscr{C}_n(\gamma)$ converges to $\gamma^\star$ almost surely, namely,*

$$|\widehat{\gamma}_n - \gamma^\star| \to 0 \ \text{ as } n \to \infty.$$

## References


Ahnert, S. E., & Payne, M. C. (2006). All possible bipartite positive-operator-value measurements of two-photon polarization states. *Physical Review A*, *73*, 022333.

Alquier, P., Butucea, C., Hebiri, M., & Meziani, K. (2013). Rank penalized estimation of a quantum system. *Physical Review A*, *88*, 032133.

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Artiles, L. M., Gill, R. D., & Guta, M. I. (2005). An invitation to quantum tomography. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *67*, 109–134.

Bisio, A., Chiribella, G., D'Ariano, G. M., Facchini, S., & Perinotti, P. (2009). Optimal quantum tomography of states, measurements, and transformations. *Physical Review Letters*, *102*, 010404.

Blume-Kohout, R. (2010). Hedged maximum likelihood quantum state estimation. *Physical Review Letters*, *105*, 200504.

Bonnabel, S., Mirrahimi, M., & Rouchon, P. (2009). Observer-based hamiltonian identification for quantum systems. *Automatica*, *45*(5), 1144–1155.

Cai, T., Kim, D., Wang, Y., Yuan, M., & Zhou, H. H. (2016). Optimal large-scale quantum state tomography with Pauli measurements. *The Annals of Statistics*, *44*(2), 682–712.

Chiuso, A. (2016). Regularization and Bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control*, *41*, 24–38.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, *57*(3), 1548–1566.

Gross, D., Liu, Y.-K., Flammia, S., & S. Becker, J. E. (2010). Quantum state tomography via compressed sensing. *Physical Review Letters*, *105*, 150401.

James, D. F. V., Kwiat, P. G., Munro, W. J., & White, A. G. (2001). Measurement of qubits. *Physical Review A*, *64*, 052312.

Leghtas, Z., Turinici, G., Rabitz, H., & Rouchon, P. (2012). Hamiltonian identification through enhanced observability utilizing quantum control. *IEEE Transactions on Automatic Control*, *57*(10), 2679–2683.

Ljung, L. (1999). *System identification: theory for the user*. Upper Saddle River, NJ: Prentice-Hall.

Nielsen, M. A., & Chuang, I. L. (2001). *Quantum computation and quantum information*. Cambridge, England: Cambridge University Press.

Petersen, K. B., & Pedersen, M. S. (2012). The matrix cookbook. http://matrixcookbook.com.

Qi, B., Hou, Z., Li, L., Dong, D., Xiang, G., & Guo, G. (2013). Quantum state tomography via linear regression estimation. *Scientific Reports*, *3*, 3496.

Qi, B., Hou, Z., Wang, Y., Dong, D., Zhong, H.-S., Li, L., et al. (2017). Adaptive quantum state tomography via linear regression estimation: theory and two-qubit experiment. *npj Quantum Information*, *3*, 19.

Rao, C. R., Toutenburg, H., Shalabh, & Heumann, C. (2008). *Linear models and generalizations: least squares and alternatives*. Berlin Heidelberg: Springer.

Recht, B., Fazel, M., & Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, *52*(3), 471–501.

Senko, C., Smith, J., Richerme, P., Lee, A., Campbell, W. C., & Monroe, C. (2014). Coherent imaging spectroscopy of a quantum many-body spin system. *Science*, *345*, 430–433.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, *4*(2), 107–194.

Smolin, J. A., Gambetta, J. M., & Smith, G. (2012). Efficient method for computing the maximum-likelihood quantum state from measurements with additive Gaussian noise. *Physical Review Letters*, *108*, 070502.

Teo, Y. S., Zhu, H., Englert, B. G., Rehacek, J., & Hradil, Z. (2011). Quantum-state reconstruction by maximizing likelihood and entropy. *Physical Review Letters*, *107*, 020404.

Theil, H. (1971). *Principles of econometrics*. John Wiley & Sons, Inc..

Wang, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *The Annals of Statistics*, *41*, 2462–2504.

Wang, Y., Dong, D., Qi, B., Zhang, J., Petersen, I. R., & Yonezawa, H. (2018). A quantum hamiltonian identification algorithm: Computational complexity and error analysis. *IEEE Transactions on Automatic Control*, *63*(5), 1388–1403.

Wang, Y., Yin, Q., Dong, D., Qi, B., Petersen, I. R., Hou, Z., et al. (2019). Quantum gate identification: Error analysis, numerical results and optical experiment. *Automatica*, *101*, 269–279.

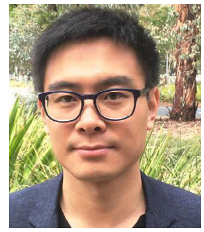
Wootters, W. K., & Fields, B. D. (1989). Optimal state-determination by mutually unbiased measurements. *Annalen der Physik*, *191*, 363–381.

Xue, S., Zhang, J., & Petersen, I. R. (2019). Identification of non-Markovian environments for spin chains. *IEEE Transactions on Control Systems Technology*, *27*(6), 2574–2580.


**Biqiang Mu** received the Bachelor of Engineering degree in Material Formation and Control Engineering from Sichuan University in 2008 and the Ph.D. degree in Operations Research and Cybernetics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2013. He is currently an assistant professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research interests include system identification (data-driven modeling and analysis), machine learning, and their applications.

**Hongsheng Qi** received his Ph.D. degree in systems theory from Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2008. From July 2008 to May 2010, he was a postdoctoral fellow in the Key Laboratory of Systems Control, Chinese Academy of Sciences. He currently is an associate professor with Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He was a recipient of "Automatica" 2008–2010 Theory/Methodology Best Paper Prize in 2011 and the second recipient of a second National Natural Science Award of China in 2014. His research interests include logical dynamic systems, game theory, quantum networks, etc.

**Ian R. Petersen** was born in Victoria, Australia. He received a Ph.D. in Electrical Engineering in 1984 from the University of Rochester. From 1983 to 1985 he was a Postdoctoral Fellow at the Australian National University. From 1985 until 2016 he was with UNSW Canberra where was most recently a Scientia Professor and an Australian Research Council Laureate Fellow in the School of Engineering and Information Technology. He has previously been ARC Executive Director for Mathematics Information and Communications, Acting Deputy Vice-Chancellor Research for UNSW and an Australian Federation Fellow. From 2017 he has been a Professor at the Australian National University. He is currently the Director of the Research School of Electrical, Energy and Materials Engineering at the Australian National University. He has served as an Associate Editor for the IEEE Transactions on Automatic Control, Systems and Control Letters, Automatica, IEEE Transactions on Control Systems Technology and SIAM Journal on Control and Optimization. Currently he is an Editor for Automatica. He is a fellow of IFAC, the IEEE and the Australian Academy of Science. His main research interests are in robust control theory, quantum control theory and stochastic control theory.



**Guodong Shi** received the B.Sc. degree in mathematics and applied mathematics from the School of Mathematics, Shandong University, Jinan, China in 2005, and the Ph.D. degree in systems theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China in 2010. From 2010 to 2014, he was a Postdoctoral Researcher at the ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden. From 2014 to 2018, he was with the Research School of Engineering, The Australian National University, Canberra, ACT, Australia, as a Lecturer and then Senior Lecturer, and a Future Engineering Research Leadership Fellow. Since 2019 he has been with the Australian Center for Field Robotics, The University of Sydney, NSW 2006, Sydney, Australia as a Senior Lecturer. His research interests include distributed control systems, quantum networking and decisions, and social opinion dynamics.