Universität zu Köln
Institut für theoretische Physik

Due to, or in spite of? The effect of constraints on efficiency in quantum estimation problems

Inaugural-Disseration
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Daniel Süß
aus Dresden

Köln, 2018

Berichterstatter (Erstgutachter):	Prof. David Gross
Berichterstatter (Zweitgutachter):	Prof. Johannes Berg
Tag der Disputation:	13. September 2018

Abstract

In this thesis, we study the interplay of constraints and complexity in quantum estimation. We investigate three inference problems, where additional structure in the form of constraints is exploited to reduce the sample and/or computational complexity. The first example is concerned with *uncertainty quantification* in quantum state estimation, where the *positive-semidefinite constraint* is used to construct more powerful, that is smaller, error regions. However, as we show in this work, doing so in an optimal way constitutes a computationally hard problem, and therefore, is intractable for larger systems. This is in stark contrast to the unconstrained version of the problem under consideration. The second inference problem deals with *phase retrieval* and its application to characterizing *linear optical circuits*. The main challenge here is the fact that the measurements are insensitive to complex phases, and hence, reconstruction requires deliberate utilization of interference. We propose a reconstruction algorithm based on ideas from *low-rank matrix recovery*. More specifically, we map the problem of reconstruction from phase-insensitive measurements to the problem of recovering a rank-one matrix from linear measurements. For the efficient solution of the latter it is crucial to exploit the rank-one constraint. In this work, we adapt existing work on phase retrieval to the specific application of characterizing linear optical devices. Furthermore, we propose a measurement ensemble tailored specifically around the limitations encountered in this application. The main contribution of this work is the proof of efficacy and efficiency of the proposed protocol. Finally, we investigate *low-rank tensor recovery* – the problem of reconstructing a low-complexity tensor embedded in an exponentially large space. We derive a sufficient condition for reconstructing low-rank tensors from product measurements, which relates the error of the initialization and concentration properties of the measurements. Furthermore, we provide evidence that this condition is satisfied with high probability by Gaussian product tensors with the number of measurements only depending on the target's intrinsic complexity, and hence, scaling polynomially in the order of tensor. Therefore, the low-rank constraint can be exploited to dramatically reduce the sample complexity of the problem. Additionally, the use of measurement tensors with an efficient representation is necessary for computational efficiency.

Kurzzusammenfassung

In dieser Arbeit untersuchen wir das Zusammenspiel von Zwangsbedingungen und Komplexität in Quantenschätzproblemen. Für diesen Zweck betrachten wir drei Schätzprobleme, deren zusätzliche Struktur in der Form von Zwangsbedingungen ausgenutzt werden kann um die notwendige Zahl der Messungen oder die Berechnungskomplexität zu verringern. Das erste Beispiel beschäftigt sich mit der *Unsicherheitsabschätzung* in *Quantenzustandstomographie*. In diesem nutzen wir die Zwangsbedingung, dass physikalische gemischte Zustände durch *positiv-semidefinite* Matrizen beschrieben werden, um kleinere und damit aussagekräftigere Fehlerregionen zu konstruieren. Wir zeigen jedoch in dieser Arbeit, dass ein optimales Nutzen der Zwangsbedingungen ein rechnerisch schweres Problem darstellt und damit nicht effizient lösbar ist. Im Vergleich dazu existieren effiziente Algorithmen für die Berechnung optimaler Fehlerregionen im Falle des Models ohne Zwangsbedingungen. Das zweite Schätzproblem beschäftigt sich mit *Phase Retrieval* und dessen Anwendung für die Charakterisierung von *linear-optischen Schaltkreisen*. Hier ist die größte Herausforderung, dass die Messungen phasenunempfindlich sind und deshalb das gezielte Ausnutzen von Interferenz für die vollständige Rekonstruktion notwendig ist. Für diesen Zweck entwickeln wir einen Algorithmus, der auf Techniken der effizienten Rekonstruktion von Matrizen mit niedrigem Rang basiert. Genauer gesagt bilden wir das Problem der Rekonstruktion aus phasenunempfindlichen Messungen auf das Problem der Rekonstruktion von Matrizen mit Rang 1 aus linearen Messungen ab. Für dessen effiziente Lösung ist es notwendig die Zwangsbedingung an den Rang auszunutzen. Hierzu adaptieren wir existierende Arbeiten aus dem Bereich des Phase Retrievals für die Anwendung auf die Charakterisierung von linear-optischen Schaltkreisen. Außerdem entwickeln wir ein Ensemble von Messvektoren, das speziell auf diese Anwendung zugeschnitten ist. Zuletzt untersuchen wir die Rekonstruktion von *Tensoren mit niedrigem Rang*, also das Problem einen Tensor mit niedriger Komplexität in einem exponentiell großem Hilbertraum zu rekonstruieren. Wir leiten eine hinreichende Bedingung für die erfolgreiche Rekonstruktion von Tensoren mit niedrigem Range aus Rang 1 Messungen ab, die den erlaubten Fehler der Initialisierung und Konzentrationseigenschaften der Messungen miteinander in Verbindung setzt. Außerdem zeigen wir numerisch, dass Gauss'sche Produktmessungen diese Eigenschaft mit hoher Wahrscheinlichkeit erfüllen, auch wenn die Zahl der Messungen polynomiell in der Ordnung des Zieltensors skaliert. Damit können die Rangbedingungen ausgenutzt werden um die Zahl der notwendigen Messungen drastisch zu

reduzieren. Zusätzlich dazu ist unser Ansatz auch recheneffizient, da wir effizient darstellbare Produkttensoren als Messungen verwenden.

Contents

1. Introduction

Physics as an inherently empirical science relies on experimental data to single out theories that are compatible with observations. It is therefore a fundamental problem to transform data into answers to questions posed by the physicist. A large fraction of experimental problems can be phrased in terms of parametric estimation: Given a mathematical model that relates the parameters of interest to observable outcomes, estimate the parameters that fit the data "best". In other words, parameter estimation is an inverse problem for a fixed model of the system.

Two crucial aspects of estimation problems in general are constraints and complexity. The former refers to the fact that many models are specified in terms of continuous parameters, but not all parameter values correspond to valid models. One typical example for constraints are mathematical facts, e.g. the standard deviation σ of a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ should be positive. Others include assumptions on the particular model such as the constraint that a valid density matrix of a quantum system is positive semi-definite. Note that not all constraints need to be hard constraints as the examples stated above. Soft constraints can be used to promote desired properties of the estimate or to penalize values due to prior knowledge.

In general, the notion of complexity refers to the amount of resources required to solve inference problems as the size of the underlying models grows. On the one hand, we use "complexity" to refer to the amount of experimental resources required, which is often phrased in terms of sample complexity, i.e. the number of measurements. Also, other measures such as the time required to perform a given experiment are possible if the necessary information can be quantified. On the other hand, we are interested in the amount of computational resources required to extract the desired information from the available data. Theses are often quantified in terms of runtime of the inference computation. Note that these two notions of complexity can be strongly related: If an experiment – such as the ones performed at the LHC – truly deserves the ubiquitous "Big Data" label, then an enormous amount of computational resources is necessary to perform even simple computations on the whole dataset. Additionally, inference problems can also be intrinsically hard to solve such as inference in general Bayesian networks [Coo90; Rot96].

In this work, we are interested in the interaction of constraints and complexity. The main motivation stems from recent progress in quantum technologies in general and quantum computing in particular: Many established techniques for characterization – i.e. inferring a complete description of a system from experimental data – work well

for small systems with only a few qubits, which were prevalent in the past. However, many of these approaches do not scale to larger systems as they require exponentially large amounts of resources. With the recent announcements of quantum devices with up to 72-qubits [Con18], it becomes clear that new techniques for characterization need to be developed. One way to go forward is to exploit physical constraints or to impose additional assumptions on the model to reduce the amount of resources necessary or to improve estimates. To better understand the interplay of constraints and complexity, we investigate three inference problems with applications to quantum estimation in this work.

In Chapter 2, we examine uncertainty quantification in quantum state estimation – the problem of estimating the density matrix ρ of a quantum system from measurements. Since all outcomes of quantum measurements are inherently random, one should not only report the final estimate, but also answer the question whether the result is statistically reliable or simply arose due to chance. For this purpose, we use the notion of "error bars" or "error regions" from statistical inference. To investigate the effects of constraints on the computational complexity of the problem, we consider models that allow for easily computable optimal error regions in the unconstrained case. For those models, we show that taking into account the physical constraints on ρ, i.e. positive semi-definiteness, renders the problem of computing optimal error regions intractable. We also show that there are settings, where exactly those physical constraints drastically improve the power of error regions, and therefore, are necessary for optimality. In conclusion, we show that exploiting the physical constraints on ρ is essential to obtain optimal error regions, but doing so in an optimal way poses an intractable computational problem in general.

The second main result of this thesis is concerned with characterizing linear optical circuits, which have been proposed as one possible architecture for quantum computing. By measuring the output of such a device for different inputs, the problem is to reconstruct the *transfer matrix* of the device. Here, the main challenge is that the standard measurement devices in optics – single photon detectors and photo diodes – are insensitive to the phases of the output. Therefore, we need to deliberately use interference between different modes of the device to gain information on the complex phases. This raises the question of how to choose the inputs in an optimal way, in order to reduce the total number of measurements required, and how to reconstruct the transfer matrix in a computationally efficient way. In Chapter 3, we propose a characterization method that is asymptotically optimal w.r.t. the sample complexity, comes with a rigorous proof of convergence, and is robust to noise. The suggested method adapts ideas from low-rank matrix recovery and leverages an exact mathematical constraint on the signal to be recovered to derive a reconstruction algorithm that is efficient w.r.t. both sample and computational complexity.

Chapter 4 is dedicated to efficiently reconstructing low-rank tensors from linear measurements. As a natural extension of low-rank matrix recovery, this challeng-

ing problem has received a lot of attention recently. Here, we consider tensors $X \in \left(\mathbb{R}^d\right)^{\otimes N}$, where d is the local dimension and N the order of the tensor. Without any additional structure, reconstructing X from measurements of the form $\langle A, X\rangle$ for measurement tensors A requires at least d^N such overlaps as each component of X is independent from the other. Therefore, the sample complexity scales exponentially in N and the problem becomes infeasible already for moderate values of N. Furthermore, any reconstruction algorithm of an arbitrary tensor requires an exponentially long runtime as simply outputting the result takes this amount of time. However, many tensors naturally occurring, e.g. in quantum physics, have additional structure that can be used to render their description and reconstruction efficient. Here, we consider low-MPS rank tensors, which are a generalization of low-rank matrices and constitute a variational class of tenors that have an efficient description in terms of the matrix product state (MPS) representation. More precisely, the number of parameters required to express a tensor of fixed MPS rank in said representation scales linearly in N. The question we are trying to answer is whether such low-MPS rank tensors can be recovered from m linear measurements such that m scales polynomially in the intrinsic complexity, i.e. polynomially in N, instead of depending on the dimension of the embedding space. Additionally, we want the reconstruction algorithm to be efficient. For this purpose, it is necessary to consider measurement tensors A that also have an efficient representation, e.g. in the MPS tensor format. In this work, we provide a partial answer to this question We derive a condition on the measurement tensors that is sufficient for recovery via an alternating-minimization algorithm. Numerically, show that random Gaussian product tensors are a viable candidate for measurement tensors fulfilling this condition. The role of the low-MPS rank constraint in this problem is in stark contrast to Chapter 2, where imposing constraints on the parameter to be inferred rendered the problem computationally intractable. Here, the problem of tensor recovery only becomes tractable with the additional low-rank constraint.

2. Uncertainty quantification for quantum state estimation

Due to the intrinsic randomness of quantum mechanics, any information we obtain about a quantum system through measurements is subject to statistical uncertainty. Consequently, one should not only report the final result of an experiment, but also answer the question whether this result is statistically reliable or simply arose due to chance. This motivated the recent development of techniques for uncertainty quantification in quantum state estimation (QSE) [Blu12; Fer14a; Sha+13; FR16; CR12; AS09; Aud+08]. Despite the large body of work concerned with this problem, none of these constructions are known to be both optimal and computationally feasible.

In this chapter, we investigate the question whether the lack of an efficient algorithm for computing optimal error regions in QSE simply is due to a lack of imagination or if there are fundamental restrictions that prevent the existence of such an algorithm. We provide evidence that the latter case is true: Based on the generally accepted conjecture that $\mathbf{P} \neq \mathbf{NP}$ in computational complexity we show that no such algorithm can exist. More specifically, we show that the computational intractability does not simply arise due to the general difficulties of statistics in high dimensions. Instead, by considering models which render the unconstrained problem tractable, we show that the computational intractability is caused by the *quantum mechanical shape constraints*: For ϱ to constitute a valid quantum state, it has to be a Hermitian, positive semi-definite (psd) matrix with unit trace. The Hermiticity and trace constraints are linear and easily satisfiable by means of a suitable linear parametrization. In contrast, the psd constraint is non-linear, and hence, more problematic to satisfy.

The motivation for studying uncertainty quantification under quantum constraints is that these constraints can lead to a significant reduction in uncertainty. This is particularly evident if the true state is close to the boundary of state space, e.g. in the case of almost pure states, which are of interest in quantum information processing experiments. In this case, it is plausible that a large fraction of possible estimates that seem compatible with the observations can be discarded, as they lie outside of state space.

Indeed, it is known that taking the quantum constraints into account can result in a dramatic – even unbounded – reduction in uncertainty. Prime examples are results that employ positivity to show that even *informationally incomplete* measurements can be used to identify a state with arbitrarily small error [Cra+10; Gro+10; Gro11;

Fla+12; NG+13; KKD15]. More precisely, these papers describe ways to rigorously bound the size of a confidence region for the quantum state based only on the observed data and on the knowledge that the data comes from measurements on a valid quantum state. While these uncertainty bounds can always be trusted without further assumptions, only in very particular situations have they been proven to actually become small. These situations include the cases where the true state is of low rank [Fla+12; NG+13], or admits an economical description as a matrix-product state [Cra+10]. It stands to reason that there are further cases – not yet identified – for which the size of an error region can be substantially reduced simply by taking into account the quantum shape constraints.

Throughout this work, we consider the task of non-adaptive QSE, where a fixed set of measurements specified by the choice of positive operator valued measure (POVM) is performed on independent copies of the system. For a given POVM, the measurement outcomes of this setup can be described in terms of a generalized linear model (GLM) with parameter ϱ – the quantum mechanical state or density matrix of the system. The well-established theory of GLMs then provides methods for inferring ϱ [MN89].

However, what sets the task of QSE apart from inference in GLMs in general are the additional quantum mechanical shape constraints. Nevertheless, there are estimators for ϱ that on the one hand always yield a psd density matrix, and on the other hand, are well-understood theoretically with near-optimal performance and scalable to intermediate sized quantum experiments [PR04]. Unfortunately, the same cannot be said for error bars for ϱ.

This chapter is structured as follows: We introduce the necessary concepts from statistical inference and computational complexity in Section 2.1 and 2.2, respectively. Then, in Section 2.3 we provide more details on QSE and the One of the main results of this work concerning the proof of computational intractability of frequentist confidence region is given in Section 2.4. The Bayesian counterpart for credible regions is the topic of Section 2.5. Finally, we conclude this chapter with remarks on limitations of the hardness results as well as possible future work in Section 2.6.

Relevant publications

- D. Suess, Ł. Rudnicki, T. O. Maciel, D. Gross: *Error regions in quantum state tomography: computational complexity caused by geometry of quantum states*, New J. Phys. 19 093013 (2017)

2.1. Introduction to statistical inference

The objective of statistical inference is to obtain information about the distribution of a random variable X from observational data. Here, we focus on the special case of inference in parametric models, which can be described as follows [Was13]: A *parametric model* is a family of distributions $\{\mathbb{P}_\theta\}_{\theta\in\Omega}$ labeled by a finite number of parameters θ, where $\Omega \subseteq \mathbb{R}^k$ is called the *state space* of the model. For simplicity, we only consider the scalar case $k = 1$ for now. Then, any function $\hat{\theta}$ mapping observations to the space $\mathbb{R}$ containing the parameter space is called a *point estimator* for the parameter θ. In general, we do not require $\hat{\theta}$ to map to the state space Ω as this restriction would preclude many relevant estimators such as the linear inversion estimator introduced in Section 2.4.2. However, since we are interested in learning about the distribution of X, not every estimator is equally useful. Indeed, the goal should be to find an estimator, which yields the parameter value that describes the observed data best. What we mean by "best" in this context not only depends on the specific model and what the estimate is supposed to be used for, but also on the fundamental interpretation of probability. Broadly speaking, there are two different interpretations of probability, namely the frequentist (or orthodox) interpretation and the Bayesian interpretation [Háj12], which lead to distinct schools of inference [Kie12; BC16; Was13]. Although the two approaches yield the same results for very simple models or – under mild regularity assumptions – in the limit of many measurements, they generally differ and in some cases even yield contradictory results [Was13, Sec. 11.9].

So far we have only discussed point estimators, which yield a single value for the parameters. However, even if our model describes the data perfectly for some choice of θ, we cannot exactly recover this value from a finite amount of data due to statistical fluctuations in general. The concept of error bars, or more specifically error regions, allows for quantifying the uncertainty of a given estimate. In Section 2.1.1, we introduce the basic concepts of frequentist inference and the corresponding notion of *confidence regions*. Section 2.1.2 is concerned with Bayesian inference and the corresponding notion of error regions, namely *credible regions*.

2.1.1. Frequentist statistics

In the frequentist framework, the probability of an outcome of a random experiment is defined in terms of its relative frequency of occurrence when the number of repetitions goes to infinity [Key07; Kie12]. More precisely, denote the number of repetitions of an experiment by N and the number of times the event under consideration x occurred by n_N. Then, a frequentist interprets the probability $\mathbb{P}(x)$ as the statement that $\frac{n_N}{N} \to \mathbb{P}(x)$ as $N \to \infty$.

For the task of parameter estimation, we assume that the model is well-specified,

i.e. that the observed data are generated from the parametric model with the fixed "true" parameter $\theta_0 \in \Omega$, which is unknown. From a finite number of observations $X_1, \dots X_N$, we must construct an estimate $\hat{\theta}(\boldsymbol{X})$ for θ_0. Although there are some intuitive approaches to this problem such as the method of moments [Was13, Sec. 9.2], the *principle of maximum likelihood* is often employed to construct an estimator with good frequentist properties. First, let us introduce the *likelihood function* of the model

$$\mathcal{L}(\theta; \boldsymbol{X}) = \mathbb{P}_\theta(\boldsymbol{X}) = \prod_{i=1}^{N} \mathbb{P}_\theta(X_i), \tag{2.1}$$

where we have assumed independence of the samples for the second equality. The *maximum likelihood estimator* (MLE) is then defined by[1]

$$\hat{\theta}_{\mathrm{MLE}}(\boldsymbol{x}) = \mathrm{argmax}_{\theta \in \Omega}\, \mathcal{L}(\theta; \boldsymbol{x}). \tag{2.2}$$

The justification for using (2.2) is that under mild conditions on the model, the MLE posses many appealing properties [Was13, Sec. 9.4] such as consistency and efficiency: Consistency means that the MLE converges to the true value θ_0 in probability as $N \to \infty$ and efficiency roughly means that among all well behaved estimators, the MLE has the smallest variance in the large sample limit.

A more flexible approach to the problem of how to single out a "good" estimator is formalized in statistical decision theory [CB02; LC98]. For this purpose, we need to introduce a *loss function*

$$\mathcal{L}\colon \Omega \times \mathbb{R} \to \mathbb{R}, (\theta_0, \hat{\theta}) \mapsto \mathcal{L}(\theta_0, \hat{\theta}), \tag{2.3}$$

which measures the discrepancy between the true value θ_0 and an estimate $\hat{\theta}(\boldsymbol{X})$. Keep in mind that since the X_i are random variables, so is $\hat{\theta}(\boldsymbol{X})$ and its loss. In order to asses the estimator $\hat{\theta}$, that is the function mapping data to an estimate, we evaluate the average loss or *risk function*

$$\mathcal{R}(\theta_0, \hat{\theta}) = \mathbb{E}_{\theta_0}(\mathcal{L}(\theta_0, \hat{\theta}(\boldsymbol{X}))) = \int \mathcal{L}(\theta_0, \hat{\theta}(\boldsymbol{x}))\, \mathbb{P}_{\theta_0}(\boldsymbol{x})\, \mathrm{d}x. \tag{2.4}$$

Then, the problem of finding a good estimator reduces to the problem of finding an estimator that yields small values for Eq. (2.4). However, note that the risk function (2.4) for a given estimator still depends on the unknown true value. To obtain a single-number summary of the performance of an estimator for all possible values of θ_0, there are different strategies such as maximizing or averaging the risk function over all θ_0 [Was13, Sec. 12.2]. By minimizing these risks over all possible estimators, we try to determine a single estimator that performs best in the worst case

[1] Note that we have constrained the values of the MLE to the state space Ω of the model. Otherwise, the right hand side of Eq. (2.2) might not be well defined.

or on average, respectively. However, in the following we introduce a less ambitious notion of optimality, namely *admissibility*.

Definition 2.1. *[Was13, Def. 12.17] A point estimator $\hat{\theta}$ is* inadmissible *if there exists another estimator $\hat{\theta}'$ such that*

$$\begin{aligned} &\mathcal{R}(\theta_0, \hat{\theta}') \leq \mathcal{R}(\theta_0, \hat{\theta}) \quad \textit{for all } \theta_0 \in \Omega \\ &\mathcal{R}(\theta_0, \hat{\theta}') < \mathcal{R}(\theta_0, \hat{\theta}) \quad \textit{for at least one } \theta_0 \in \Omega. \end{aligned}$$

Otherwise, we call $\hat{\theta}$ admissible.

In words, $\hat{\theta}$ is admissible if there is no other estimator $\hat{\theta}'$ that performs at least as good as $\hat{\theta}$ and strictly better for at least one value of the true value θ_0.

The choice of loss function generally depends on the problem at hand and determines the properties of the corresponding estimator. Some examples for commonly used loss functions include the 0/1-loss for discrete parameter models

$$\mathcal{L}(\theta_0, \theta) = \begin{cases} 1 & \theta_0 = \theta \\ 0 & \text{otherwise} \end{cases}, \tag{2.5}$$

and the mean squared error (MSE)

$$\mathcal{L}(\theta_0, \theta) = (\theta_0 - \theta)^2, \tag{2.6}$$

which is often used for continuous parameter models, e.g. $\Omega = \mathbb{R}$. The use of the MSE loss is often motivated by the fact that it gives the same results as the principle of maximum likelihood for many problems. As an example, consider the task of estimating the mean from iid normal random variables $X^{(k)} \sim \mathcal{N}(\theta_0, \sigma)$ with $\sigma \in \mathbb{R}_+$ known. In this case, a straightforward estimator for θ_0 is the *empirical mean*

$$\bar{X} := \frac{1}{m} \sum_{i=1}^{m} X^{(k)}, \tag{2.7}$$

which is admissible with respect to MSE [Was13, Thm. 12.20]. Furthermore, $\bar{X}$ is also the MLE as a transformation to log-likelihood shows:

$$\begin{aligned} \operatorname{argmax}_\theta \mathcal{L}(\theta; x^{(1)}, \dots, x^{(m)}) &= \operatorname{argmax}_\theta \sum_{k=1}^{m} \log \mathbb{P}_\theta(X^{(k)} = x^{(k)}) \\ &= \operatorname{argmin}_\theta \sum_{k=1}^{m} \frac{1}{2\sigma^2} \Big(x^{(k)} - \theta \Big)^2, \end{aligned} \tag{2.8}$$

where we have used that the $X^{(k)}$ are independent and that the logarithm is monotonic increasing. Furthermore, we discarded all contributions independent of θ is the

second line. Finally, note that the right hand side of the last equation is minimized by the choice $\theta = \bar{x}$, which proofs the claim.

Now we consider a multivariate generalization of the above problem, that is the task of estimating $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ from a linear Gaussian model $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}_0, \mathbb{1})$ with unit covariance matrix. Suppose we only have a single observation $\boldsymbol{X}$, i.e. $m = 1$. Since all the components X_i are independent, one expects that a separate estimation of the components via $\bar{\boldsymbol{X}} = \boldsymbol{X}$ constitutes a "good estimator". Indeed, the same computation as in Eq. (2.8) shows that the empirical mean is the MLE for this model as well. However, Stein shocked the statistical community when he proved that for $d \geq 3$, this estimator is inadmissible [Ste+56]. It can be shown that the *James-Stein* estimator

$$\hat{\boldsymbol{\theta}}_{\mathrm{S}} := \max\left\{0, \left(1 - \frac{d-2}{\|\boldsymbol{X}\|_2}\right)\right\} \boldsymbol{X} \tag{2.9}$$

has smaller MSE risk than the empirical mean for all values of $\boldsymbol{\theta}_0$ [Ste+56; LC98]. It is often referred to as a *shrinkage estimator* because it shrinks the empirical mean estimate $\boldsymbol{X}$ towards 0. Note however, that the choice of the origin as the fix point of the shrinkage operation is arbitrary – the James-Stein estimator outperforms the empirical mean estimator for any choice of fix point. Interestingly, Eq. (2.9) shows that by taking into account all the components X_i, we can improve the MSE of the mean-estimator even if the X_i are independent. However, this only applies to the *simultaneous* MSE error of all the components of $\boldsymbol{\theta}_0$. The James-Stein estimator (2.9) cannot be used to improve the MSE of a single component. Equation (2.9) can also be generalized to the case of more than one observation, i.e. $m > 1$. Finally, note that the James-Stein estimator is also not admissible: More elaborate shrinkage estimators are known to outperform (2.9) w.r.t. the MSE. Even worse, to the best of the author's knowledge, no admissible construction exists for estimating the mean of a d-variate Gaussian w.r.t. MSE for $d \geq 3$.

As already mentioned in the introduction, point estimators cannot convey uncertainty in the estimate. For this purpose, we need to introduce a precise notion of "error bars", namely *confidence regions* in the framework of frequentist statistics.

Definition 2.2. *Consider a statistical model with k parameters, that is $\Omega \subseteq \mathbb{R}^k$. A confidence region $\mathcal{C}$ with coverage $\alpha \in [0,1]$ is a region estimator – that is a function that maps observed data x to a subset $\mathcal{C}(x) \subseteq \mathbb{R}^k$ of the space containing the state space – such that the true parameter is contained in $\mathcal{C}$ with probability greater than α:*

$$\forall \theta_0 \in \Omega \colon \mathbb{P}_{\theta_0}\left(\mathcal{C}(X) \ni \theta_0\right) \geq 1 - \alpha. \tag{2.10}$$

Note that the coverage condition (2.10) is not a probabilistic statement about the true parameter θ_0 for fixed observed data x. Instead, Definition 2.2 should

be interpreted as a statistical guarantee for the region estimator $\mathcal{C}$: Say we repeat an experiment m times and obtain data $x^{(1)}$, ...$x^{(M)}$ from a distribution with true parameter θ_0. Then, in the limit $m \to \infty$ at least a fraction of $1-\alpha$ of the regions $\mathcal{C}(x^{(1)})$, ..., $\mathcal{C}(x^{(m)})$ contain the true parameter θ_0. In other words, the probabilistic statement in (2.10) refers to the random variable $\mathcal{C}(X)$ for fixed θ_0.

Similarly to point estimators, Eq. (2.10) does not uniquely determine a confidence region construction. Additional constraints are necessary to exclude trivial constructions such as the following: Take the region estimator that is always equal to the full parameter space independent of the data $\mathcal{C}(X) = \Omega$, then

$$\mathbb{P}_\theta\left(\mathcal{C}(X) \ni \theta_0\right) = 1 \geq 1 - \alpha \tag{2.11}$$

for all confidence levels α. Although, this construction trivially fulfils the coverage condition (2.10), it does not provide useful information on the uncertainty as it does not restrict the parameter space at all. Therefore, we have to impose a notion of what constitutes a good confidence region by introducing a loss function. Clearly, if we have two confidence regions $\mathcal{C}_1$ and $\mathcal{C}_2$ with the same confidence level and $\mathcal{C}_1 \subset \mathcal{C}_2$, then $\mathcal{C}_1$ is more informative. More generally, smaller regions should be preferred since they convey more confidence in the estimate and exclude more alternatives. Therefore, measures of size such as (expected) volume or diameter are commonly used as loss functions for region estimators. We now introduce a notion of optimality similar to Definition 2.1.

Definition 2.3. *[Jos69, Def. 2.2] A confidence region $\mathcal{C}$ for the parameter estimation of $\theta_0 \in \Omega$ is called (weakly)* admissible *if there is no other confidence region $\mathcal{C}'$ that fulfils*

1. *(equal or smaller volume) $\mathcal{V}(\mathcal{C}'(x)) \leq \mathcal{V}(\mathcal{C}(x))$ for almost all observations x.*
2. *(same or better coverage) $\mathbb{P}_{\theta_0}(\mathcal{C}' \ni \theta_0) \geq \mathbb{P}_{\theta_0}(\mathcal{C} \ni \theta_0)$ for all $\theta_0 \in \Omega$.*
3. *(strictly better) strict inequality holds for one $\theta_0 \in \Omega$ in (ii) or on a set of positive measure in (i).*

This notion of optimality uses "pointwise" volume as a risk function instead of the average volume. The conditions in Definition 2.3 are stated only for "almost all" x since one can always modify the region estimators on sets of measure zero without changing their statistical performance. In Section 2.4.1 we show that this notion of admissibility is especially suitable for studying constrained inference problems.

2.1.2. Bayesian statistics

Let us now introduce the Bayesian point of view on statistical inference: In the Bayesian interpretation, probabilities do not describe frequencies in the limit of infinitely many repetitions, but they reflect subjective degrees of belief. Put differently,

in the Bayesian framework randomness reflects one's ignorance or lack of knowledge of the value of a parameter. In contrast to frequentist inference, this enables us to make probabilistic statements about the values of parameters even for a single fixed set of observations. Generally, Bayesian inference for a parametric model is carried out in the following way: First, we choose a *prior distribution* $\mathbb{P}(\theta)$, which expresses our belief about the parameter θ before taking any data into account. Given an observation X, the distribution of θ is updated according to Bayes' rule [BC16; Gel+95]

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}. \tag{2.12}$$

Here, $\mathbb{P}(X|\theta)$ is the *likelihood function* of the model analogous to (2.1) and $\mathbb{P}(\theta|X)$ is the *posterior distribution* – or short posterior – of θ. Of course, the update in Eq. (2.12) is not limited to a single observation and can be iterated for independent data.

Computing the Bayesian update (2.12) analytically is possible only in a few rare cases: If for a given likelihood function, the prior and the posterior are in the same family of distributions, the prior is called a *conjugate prior* for the likelihood function. For example, consider a Gaussian random variable $X \sim \mathcal{N}(\theta, \tau^2)$ with known variance τ^2. If we assume a Gaussian prior $\theta \sim \mathcal{N}(\mu, \sigma^2)$, we have [Gel+95, Eq. (2.10)]

$$\theta|X = x \sim \mathcal{N}(\mu', \sigma'^2) \tag{2.13}$$

with

$$\mu' = \frac{\frac{\mu}{\sigma^2} + \frac{x}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \quad \text{and} \quad \sigma' = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-\frac{1}{2}}. \tag{2.14}$$

In other words, an update of a Gaussian prior with a Gaussian likelihood function yields a Gaussian posterior distribution. Although there are other well-known conjugate priors with explicit formulas for the parameter update, in practice the Bayesian update (2.12) can only be approximated numerically. Commonly used methods include sampling techniques such as Markov Chain Monte Carlo and Sequential Monte Carlo [Gel+95] as well as variational Bayes [FR12].

Note that the posterior distribution encodes all the information on θ we have. To summarize important features of the posterior, we introduce point and region estimators similar to the frequentist case: One commonly used estimator is the Bayesian mean estimator (BME) given by

$$\hat{\theta}_{\mathrm{BME}}(X) = \int \theta\, \mathbb{P}(\theta|X)\, \mathrm{d}\theta. \tag{2.15}$$

A justification for the BME is that it minimizes (under normal circumstances [LC98; BGW05]) the expected risk $\mathbb{E}\left(\mathcal{R}(\theta, \hat{\theta}(X))\right)$, provided the corresponding loss function $\mathcal{L}$ is a *Bregman divergence*. Note that the expectation is taken over θ w.r.t. the

posterior with observation X. Another example of a point estimator is the *maximum a posteriori* (MAP) estimator. As the name suggests, the MAP estimator is obtained by maximizing the posterior (2.12). Since this does not require the expensive computation of the denominator in Eq. (2.12) and – at least local – minima of the posterior can be found efficiently via gradient descent, the MAP estimator is often used for high dimensional problems in machine learning [Mur12] that otherwise would be intractable.

Let us now introduce the appropriate concept of error regions in the Bayesian framework.

Definition 2.4. *A* credible region[2] $\mathcal{C}$ *with credibility* $1-\alpha$ *is a subset of the parameter space* Ω *containing at least mass* $1-\alpha$ *of the posterior*

$$\mathbb{P}(\theta \in \mathcal{C} | X_1, \dots, X_N) \geq 1 - \alpha. \tag{2.16}$$

Notice the different notion of randomness compared to Eq. (2.10): Confidence regions are random variables due to their dependence on the data and Eq. (2.10) demands that the true value is contained in the confidence region with high probability. Here, "probability" refers to (possibly hypothetical) repetitions of the experiment – no statement can be made for a single run of an experiment with given outcomes. In contrast, the definition of credible regions (2.16) only refers to probability w.r.t. the posterior, and therefore, is well defined even for a single run of the experiment.

In order to single out "good" credible regions, we need to introduce a notion of optimality. As argued in Section 2.1.1, smaller error regions are generally more informative. Therefore, good credible regions are minimal-volume credible regions (MVCRs) or credible regions with the smallest diameter. Although in the following we deal with MVCRs w.r.t. a geometric notion of volume, some authors have proposed to measure the volume of a region by its prior probability [EGS06; Sha+13]. In case the posterior has probability density Π w.r.t. the volume measure under consideration, the MVCR is given by highest posterior sets [Fer14a]

$$\mathcal{C} = \{\theta \in \Omega \colon \Pi(\theta) \geq \lambda\}. \tag{2.17}$$

The constant λ is determined by the saturation of the credibility level condition (2.16).

2.2. Introduction to computational complexity theory

The objective of computational complexity theory is to classify computational problems according to their inherent difficulty. Broadly speaking, we quantify the difficulty by the runtime needed to solve the problem. Other resources, which one may

[2] We use the same letter for confidence and credible regions when the meaning is clear from the context.

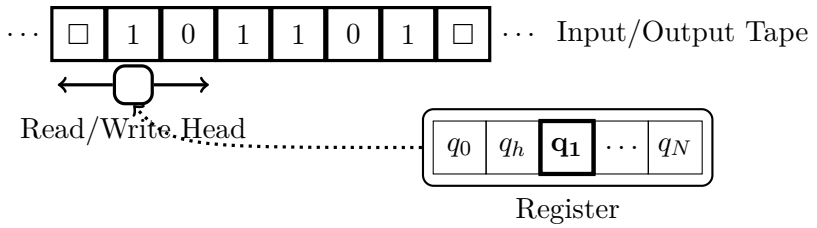


Figure 2.1.: Illustration of a Turing machine with alphabet $\Gamma = \{0, 1, \square\}$ and register $Q = \{q_0, q_h, q_1, \ldots, q_N\}$.

want to consider, are the amount of memory or communication between parties required.

At a first glance, an answer to the question whether a given problem can be solved efficiently might depend on what we consider as a "computer": Clearly, a modern high-performance cluster can solve problems in seconds, which would take a human equipped with pen and paper more than a lifetime to finish. Maybe surprisingly, it turns out that a single, simple mathematical model – the *Turing machine* – describes the capabilities and restrictions of almost all physical implementations of computation well enough for the purpose of complexity theory. The only conceivable exception so far are quantum computers that exploit quantum mechanical phenomena. Although so far no unconditional proof of an advantage of quantum computers exists, there is overwhelming evidence that they are able to solve problems efficiently for which there is no efficient classical algorithm. Note that quantum computers only provide efficient algorithms for problems that are considered hard classically. The class of uncomputable functions is the same for classical and quantum computers [AB09].

This section introduces the main concepts needed for the hardness results of Section 2.4 and 2.5. Especially the definition of **NP**-hard computational problems and polynomial-time reductions is crucial for the rest of this chapter. For a more thorough treatment, we refer the reader to [AB09; GJ79].

A Turing machine (TM) can be thought of as a simplified and idealized mathematical model of an electronic computer. It is defined by a tuple (Γ, Q, δ) and figuratively speaking consists of the following parts as shown in Fig. 2.1:

- An infinite tape, that is a bi-directional line of cells that can take the values from a finite set Γ called the *alphabet*. Γ must contain a designated symbol $\square$ called "blank".

- A register that can take on values from a finite set of states Q. Q must contain the initial state q_0 and the halting state q_h.

- A transition function

$$\delta\colon Q \times \Gamma \to Q \times \Gamma \times \{L, S, R\}, \tag{2.18}$$

  which describes the "programming" of the TM.

The operation of a TM can be summarized as follows. Initially, the reading head of the tape is over a certain cell and the register is in the initial state q_0. Furthermore, we assume that only a finite number of tape cells have a value different from $\square$ – these are referred to as the input. For one step of the computation, denote by $\gamma \in \Gamma$ the value of tape cell under the reading head and by $q \in Q$ the current value of the register. The action of the TM is determined by the transition function

$$(q', \gamma', h) = \delta(q, \gamma). \tag{2.19}$$

as follows: The register is set to value q', the tape head overwrites the cell with symbol γ', and it moves depending on the value of h: If $h = L$ or $h = R$, it moves one cell to the left or right, respectively and if $h = S$ it stays in the current position. This cycle repeats until the register takes on the halting state q_h. If the TM halts, the state of the tape with leading and trailing blanks removed is taken as its output. Note that the definition of a TM given above is one of many; others may include additional scratch tapes or have tapes that only extend to infinity only in one direction [AB09]. However, they are all equivalent, i.e. the architecture given above can simulate all the other architectures with a small overhead and vice versa. Furthermore, we can assume $\Gamma = \{0, 1, \square\}$, that is, we use a binary encoding for all non-blank values of the alphabet.

We now formalize the notion of runtime of a TM and the complexity of a given problem. Denote by

$$\{0,1\}^* = \bigcup_{n \in \mathbb{N}} \{0,1\}^n \tag{2.20}$$

the set of all finite bit strings.

Definition 2.5. *[AB09, Def. 1.3] Let $f\colon \{0,1\}^* \to \{0,1\}^*$ and $T\colon \mathbb{N} \to \mathbb{N}$ be some functions. Furthermore, let M be a Turing machine. We say M* computes *f in $T(n)$-time if for every input $x \in \{0,1\}^*$, whenever M is initialized with input x, it halts with $f(x)$ written on its output tape in at most $T(|x|)$ steps. Here, $|x|$ denotes the length of x.*

In words, Definition 2.5 gives a formal notion of the runtime of a TM for computing a function f in terms of the number of elemental steps required. But as stated in the introduction, we are more interested in the intrinsic complexity of computing the function f instead of a specific implementation. For this purpose, we introduce *complexity classes*, which are sets of functions that can be computed within given

resource bounds. The most important examples of complexity classes pertain to boolean functions $f\colon \{0,1\}^* \to \{0,1\}$, which correspond to decision problems. The corresponding set of "truthy" inputs $L = \{x\ in \{0,1\}^*\colon f(x) = 1\}$ is referred to as a *language*.

Definition 2.6. *Let $T\colon \mathbb{N} \to \mathbb{N}$ be some function. We say that a language L is in* $\mathbf{DTIME}(T(n))$ *if and only if there is a TM M that computes f in time $c \times T(n)$ for some constant $c > 0$. The complexity class* $\mathbf{P}$ *is then defined by*

$$\mathbf{P} = \bigcup_{\lambda \geq 1} \mathbf{DTIME}(n^\lambda). \tag{2.21}$$

In words, $\mathbf{P}$ is the set of all languages that can be decided by a TM in a number of steps that scales polynomially in the input size. The problems in $\mathbf{P}$ are considered to be efficiently solvable. Therefore, in order to show that some problem is "easy", we just have to provide a TM, or put differently an algorithm, that decides the problem in polynomial time. On the other hand, showing that no polynomial-time algorithm exists for a given problem shows that it is computationally hard. However, proving the nonexistence of efficient algorithms for many natural computational problems has turned out to be a tremendous challenge – notwithstanding deep results for computational models strictly less powerful than the TM model [AB09, Part Two]. Hence, we are going to follow a less ambitious, but very fruitful strategy: Instead of proving that a given problem is infeasible to solve efficiently, we are comparing its computational complexity to other problems that are conjectured to be computationally infeasible. If we can show that the problem under consideration is at least as hard as many other problems, which could not be solved efficiently by a myriad of computer scientists in the last decades, then we have strong evidence to believe it is intrinsically hard. This idea is formalized in the definition of polynomial-time reductions.

Definition 2.7. *Def. 2.2 from [AB09] A language $L \subseteq \{0,1\}^*$ is polynomial-time (Karp) reducible to a Language $L' \subseteq \{0,1\}^*$ denoted by $L \leq_p L'$, if there is a polynomial-time computable function $f\colon \{0,1\}^* \to \{0,1\}^*$ such that for all $x \in \{0,1\}^*$ we have $x \in L \iff f(x) \in L'$.*

Less formally speaking, if we have $L \leq_p L'$ then L' is at least as hard to decide as L. Indeed, by using the reduction f, we can turn any TM deciding L' into a TM deciding L with at most polynomial runtime overhead. Particularly, if additionally $L' \in \mathbf{P}$ then so is L. Conversely, if no efficient algorithm for L exists and $L \leq_p L'$, then there cannot be an efficient algorithm L'. The latter observation is the basis of the strategy mentioned above: We show that a given problem is computationally hard by establishing that is at least as hard as a large class of other problems, which have withstood numerous attempts at solving them efficiently so far. For this purpose we introduce the complexity class $\mathbf{NP}$.

Definition 2.8. *A language $L \subseteq \{0,1\}^*$ is in* **NP** *if there exists a polynomial-time computable function M such that for every $x \in \{0,1\}^*$*

$$x \in L \iff \exists u \in \{0,1\}^{p(|x|)} s.t.\ M(x,u) = 1. \tag{2.22}$$

The function M is called the verifier and for $x \in L$, the bitstring u is called a certificate for x.

Clearly, $\mathbf{P} \subseteq \mathbf{NP}$ with $p(|x|) = 0$. On the other hand, the question whether $\mathbf{NP} \subseteq \mathbf{P}$, and hence, $\mathbf{P} = \mathbf{NP}$ is one of the major unsolved problems in math and science [Coo06; Aar; GJ79]. One argument against this hypothesis goes as follows: Whereas the problems in **P** are considered to be easily solvable, the problems in **NP** are at least easily checkable given the certificate. In other words the question whether $\mathbf{P} = \mathbf{NP}$ boils down to the question whether finding the solution to a problem is harder than verifying whether a given solution is correct. However, these philosophical considerations are not the main reason for the importance of Definition 2.8 from a computer scientific point of view. The main justification of the class **NP** is twofold: On the one hand, there are a myriad of problems known to be in **NP** and many of them have resisted all efforts of finding a polynomial-time algorithm. On the other hand, there is a tremendous number of problems known to be at least as hard as any problem in **NP** [GJ79] – these are referred to as **NP**-hard problems:

Definition 2.9. *We say that a language L is* **NP***-hard if for every $L' \in \mathbf{NP}$ we have $L' \leq_p L$. We say that a language L is* **NP***-complete if $L \in \mathbf{NP}$ and L is* **NP***-complete.*

Clearly, for any **NP**-hard problem L, $L \leq_p L'$ implies that L' is **NP**-hard as well, and hence, any problem in **NP** is polynomial-time reducible to L'. Therefore, an efficient algorithm for L' would provide an efficient algorithm for a large number of other problems, which are widely considered difficult and which have been confounding experts for years. This fact is taken as very strong evidence that L' cannot be solved efficiently. As an example of an **NP**-complete problem, we consider the *number partition* problem.

Problem 2.10. *Given a vector $\boldsymbol{a} \in \mathbb{N}^d$, decide whether there exists a vector $\boldsymbol{\psi}$ with*

$$\forall k\ \psi_k \in \{-1,1\} \quad \textit{and} \quad \boldsymbol{a} \cdot \boldsymbol{\psi} = 0. \tag{2.23}$$

In case there is such a vector $\boldsymbol{\psi}$ one says that the instance $\boldsymbol{a}$ allows for a partition because the sum of components of $\boldsymbol{a}$ labeled by $\psi_i = 1$ is equal to the sum of components a_i labeled by $\psi_i = -1$. For a proof of **NP**-hardness of Problem 2.10, see [GJ79].

The question remains what the notion of **NP**-hardness means in practice. Due to its strict definition in terms of worst-case runtime needed for any instance, Definition 2.8 leaves open many possibilities for the existence of "good-enough" solutions. First, approximative or probabilistic algorithms often suffice for all practical purposes and are not bounded by **NP**-hardness results [GJ79; AB09]. A prime example is the number partition problem 2.10: It is often considered the "easiest hard problem" because although it is **NP**-hard, highly efficient approximative algorithms for it exist [Kel+03]. Furthermore, considering only the worst-case behavior is often too pessimistic in practice. A more appropriate classification of the difficulty of "typical" instances is given in terms of *average case* complexity [AB09].

2.3. Introduction to QSE

The goal of state estimation is to provide a complete description of an experimental preparation procedure from experimental feasible measurements. In the case of QSE, this complete description is given in terms of the density matrix ϱ of the system [PR04]. Another common procedure performed in quantum experiments is *quantum process estimation*, where the goal is to recover a *quantum channel* [NC10]. However, the task of process estimation can be mapped to QSE by way of the Choi-Jamiolkowski isomorphism [NC10; JFH03; Alt+03], and therefore, we are only concerned with the problem of QSE in this work.

Since its inception in the fifties [Fan57], QSE has proven to be a crucial experimental tool – in particular in quantum information-inspired setups. It has been used to characterize quantum states in a large number of different platforms [OBr+04; Lun+09; Mol+04; KRB08; Rip+08; Ste+06; Chi+06; Rie+07; Sch+14] and even been scaled to systems with dimension on the order of 100 [Häf+05]. However, as the Hilbert space dimensions of quantum systems implemented in the lab grow, it is unclear whether this approach to "quantum characterization" will continue to make sense.

Acquiring and post-processing the data necessary for fully-fledged QSE is already prohibitively costly for intermediated sized quantum experiments. As an extreme example, the eight qubit MLE reconstruction with bootstrapped error bars in [Häf+05] required weeks of computation in 2005 (private communication, see [Gro+10]). Some sol This problem can be partially mitigated by means of more efficient algorihtms [SGS12; Qi+13; Hou+16; SZN17] or approaches exploiting structural assumptions to improve the sampling and computational complexity of state estimation [Cra+10; Gro+10; Fla+12; Sch+15; Bau+13b; Bau+13a].

There is also the question what use is a giant density matrix with millions of entries to an experimentalist? In many cases, the full quantum state contains more information than necessary. For example, consider the case when a single number such

as the fidelity of the state in the lab w.r.t. some target state is sufficient whether the experiment works "sufficiently well" or not. For this purpose, a variety of theoretical tools for *quantum hypothesis testing*, *certification*, and scalar *quantum parameter estimation* [OW15; Aud+08; GT09; FL11; Sch+15; Li+16] have been developed in the past years that avoid the costly step of full QSE.

However, there remain use cases that necessitate fully-fledged QSE. We see a particularly important role in the emergent field of *quantum technologies*: Any technology requires means of certifying that components function as intended and, should they fail to do so, identify the way in which they deviate from the specification. As an example, consider the implementation of a quantum gate that is designed to act as a component of a universal quantum computing setup. One could use a certification procedure – direct fidelity estimation, say – to verify that the implementation is sufficiently close to the theoretical target that it meets the stringent demands of the quantum error correction threshold. If it does, the need for QSE has been averted. However, should it fail this test, the certification methods give no indication *in which way* it deviated from the intended behavior. They yield no actionable information that could be used to adjust the preparation procedure. The pertinent question "what went wrong" cannot be cast as a hypothesis test. Thus, while many estimation and certification schemes can – and should – be formulated without resorting to full state estimation, the above example shows that QSE remains an important primitive.

2.3.1. Existing work on error regions

In practice (e.g. [Häf+05]), uncertainty quantification for tomography experiments is usually based on general-purpose resampling techniques such as "bootstrapping" [ET94]. A common procedure is this: For every fixed measurement setting, several repeated experiments are performed. This gives rise to an empirical distribution of outcomes for this particular setting. One then creates a number of simulated data sets by sampling randomly from a multinomial distribution with parameters given by the empirical values. Each simulated data set is mapped to a quantum state using maximum likelihood estimation. The variation between these reconstructions is then reported as the uncertainty region. There is no indication that this procedure grossly misrepresents the actual statistical fluctuations. However, it seems fair to say that its behavior is not well-understood. Indeed, it is simple to come up with pathological cases in which the method would be hopelessly optimistic: E.g. one could estimate the quantum state by performing only one repetition each, but for a large number of randomly chosen settings. The above method would then spuriously find a variance of zero.

On the theoretical side, some techniques to compute rigorously defined error bars

for quantum tomographic experiments have been proposed in recent years. The works of Blume-Kohout [Blu12] as well as Christandl, Renner, and Faist [CR12; FR16] exhibit methods for constructing confidence regions for QST based on likelihood level sets. While very general, neither paper provides a method that has both a runtime guarantee and also adheres to some notion of non-asymptotic optimality [Kie12; Le 12].

Some authors have proposed a "sample-splitting" approach, where the first part of the data is used to construct an estimate of the true state, whereas the second part serves to construct an error region around it [Fla+12] (based on [FL11]), as well as [Car+15b]. These approaches are efficient, but rely on specific measurement ensembles (operator bases with low operator norm), approach optimality only up to poly-logarithmic factors, and – in the case of [Fla+12; FL11] – rely on adaptive measurements.

Regarding Bayesian methods, the *Kalman filtering* techniques of [Aud+08] provide a efficient algorithm for computing credible regions. This is achieved by approximating all Bayesian distributions over density matrices by Gaussians and restricting attention to ellipsoidal credible regions. The authors develop a heuristic method for taking positivity constraints into account – but the degree to which the resulting construction deviates from being optimal remains unknown. A series of recent papers aim to improve this construction by employing the *particle filter* method for Bayesian estimation and uncertainty quantification [GFF17; Wie+15; Fer14a]. Here, Bayesian distributions are approximated as superpositions of delta distributions and credible regions constructed using Monte Carlo sampling. These methods lead to fast algorithms and are more flexible than Kalman filters with regard to modelling prior distributions that may not be well-approximated by any Gaussian. However, once more, there seems to be no rigorous estimate for how far the estimated credible regions deviate from optimality. Finally, the work in [Sha+13] constructs optimal credible regions w.r.t. a different notion of optimality: Instead of penalizing sets with larger volume, they aim to minimize the prior probability as suggested by [EGS06].

2.3.2. The QSE statistical model

We now introduce the statistical model and the corresponding likelihood function used for the rest of this chapter. The first major assumption is that the system's state in the lab is sufficiently stable for the duration of the experiment. Therefore, we can assume that all data is generated from a fixed, but unknown state $\varrho_0 \in \mathcal{S}$, where

$$\mathcal{S} := \left\{ \varrho \in \mathbb{C}^{d\times d} \colon \varrho^\dagger = \varrho, \operatorname{tr} \varrho = 1, \varrho \geq 0 \right\} \tag{2.24}$$

denotes the state space of density matrices. Note that this assumption is not necessary for QSE in general, see e.g. [GFF17] for Bayesian methods that allow for tracking

time-dependent states.

Since we consider the case of non-adaptive state estimation, the measurements performed are characterized by a fixed POVM $\{E_k\}_{k=1}^m$ and the probability of the event k when the system is in the state ϱ_0 is given by the *Born rule*

$$p_k = \operatorname{tr} E_k \varrho_0. \tag{2.25}$$

However, in reality the quantum expectation values are never observed directly. Instead, when the experiment is repeated on N independent copies of ϱ_0, the observations are counts $n_i \geq 0$ with $\sum_i n_i = N$ following a multinomial distribution

$$\mathbb{P}_{\boldsymbol{p}}(n_1, \ldots, n_m) = \frac{N!}{n_1! \cdots n_m!} p_1^{n_1} \times \cdots \times p_m^{n_m}. \tag{2.26}$$

In case N is large and all the p_i are sufficiently large, the multinomial distribution (2.26) is local asymptotic normal [Sev05]. Therefore, we can approximate Eq. (2.26) by a Gaussian distribution

$$\mathbb{P}_{\boldsymbol{p}}(y_1, \ldots, y_m) \approx \Pi_{\boldsymbol{p},\Sigma}(y_1, \ldots, y_m) \tag{2.27}$$

with $y_i = \frac{n_i}{N}$ and $\Pi_{\boldsymbol{p},\Sigma}$ denoting the probability density of a Gaussian distribution with mean $\boldsymbol{p}$ and covariance matrix $\Sigma = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T$. Hence, under this Gaussian approximation, the relative counts y_i are given in terms of a *linear Gaussian model* with Gaussian likelihood function

$$\mathcal{L}(\varrho; \boldsymbol{y}) = \mathbb{P}_{\boldsymbol{p}}(\boldsymbol{y}) \tag{2.28}$$

defined in terms of Eq. (2.27) and the probabilities $p_k = \operatorname{tr} E_k \varrho$. However, if some of the p_i are close or equal to zero, which happens for rank deficient ϱ_0, local asymtotic normality (LAN) – that is the approximation in Eq. (2.27) – does not hold. In [SB18] the authors discuss some implications of the lack of LAN in QSE.

2.4. Hardness results for confidence regions

In this section we are going to present the first major result of our work [SRG+17] concerned with frequentist confidence regions in QSE. Optimal confidence regions for high-dimensional parameter estimation problems in are generally intricate even without additional constraints as there are only few elementary settings, where optimal confidence regions are known and easily characterized.

Since the goal of this work is to demonstrate that quantum shape constraints severely complicate even "classically" simple confidence regions, in the further discussion we restrict the discussion to a simplified setting: We focus on confidence ellipsoids for Gaussian distributions, which are one of the few easily characterizable

examples. Furthermore, Gaussian distributions arise in the limit of many measurements due to *local asymptotic normality*. In this section we show that even characterizing these highly simplifying ellipsoids with the quantum constraints taken into account constitutes a hard computational problem. This simplification is motivated by the goal to show that the computational intractability exclusively stems from the quantum constraints and that it is not caused by difficulties of high-dimensional statistics in general. Furthermore, any less restrictive formulation encompassing this simplified setting must be at least as hard to solve.

In conclusion, although exploiting the physical constraints may help to reduce the uncertainty tremendously as mentioned in the introduction, doing so in an optimal way is computationally intractable in general. Therefore, our work can be interpreted as a trade-off between computational efficiency and statistical optimality in QSE.

2.4.1. Optimal confidence regions for quantum states

As already indicated in the introduction, additional constraints on the parameter θ_0 under consideration can be exploited to possibly improve any confidence region estimator. This is especially clear for notions of optimality with a loss function stated in terms of a volume measure $\mathcal{V}(\cdot)$, as we will show in this section. Therefore, assume that $\theta_0 \in \Omega_c$, where the constrained parameter space $\Omega_c \subseteq \Omega$ has non-zero measure w.r.t. $\mathcal{V}$. We consider an especially simple procedure to take the constraints into account, namely truncating all $\theta \notin \Omega_c$ from tractable confidence regions for the unconstrained problem.

Although such an ad-hoc approach does not seem to exploit the constraints in an optimal way, it has multiple advantages as we discuss in more detail in Section 2.4.3: First and foremost, some notions of optimality, e.g. admissibility, are preserved under truncation as shown in Lemma 2.11. In other words, there are notions of optimality such that truncation of an optimal confidence region for the unconstrained problem gives rise to an optimal region for the constrained one. Furthermore, as already mentioned in the introduction, our goal is to show that the intractability arises purely due taking the constraints imposed by quantum mechanics into account. Therefore, we start from a tractable solution in the unconstrained setting and show that even this simple approach of taking the constraints into account leads to an computationally intractable problem. Finally, the truncation approach simplifies the discussion but as we discuss in Section 2.4.3, our results apply to a much larger class of confidence regions such as likelihood ratio based Gaussian ellipsoids.

We start by showing that the truncation of confidence regions preservers admissibility. Notice that Definition 2.3 can be stated for both, the unconstrained estimation problem $\theta_0 \in \Omega$ as well as the constrained estimation problem $\theta_0 \in \Omega_c$. The question is: How are admissible confidence regions for the constrained setting related to

admissible confidence regions for the unconstrained estimation problem?

Lemma 2.11. *Let $\mathcal{C}$ denote an admissible confidence region for the unconstrained estimation problem for the parameter $\theta_0 \in \Omega$. Then, the truncated region estimator $\mathcal{C}^\cap := \mathcal{C} \cap \Omega_\mathrm{c}$ is an admissible confidence region for the constrained problem with $\theta_0 \in \Omega_\mathrm{c}$.*

Proof. Under the assumption that $\mathcal{C}^\cap$ is not admissible, there must exist a better confidence region $\mathcal{C}^+$ for the constrained parameter estimation problem. W.l.o.g. assume that both $\mathcal{C}^+$ and $\mathcal{C}^\cap$ have the same coverage. Therefore, we must have $\mathcal{V}(\mathcal{C}^+(\boldsymbol{y})) \leq \mathcal{V}(\mathcal{C}^\cap(\boldsymbol{y}))$ for almost all observations $\boldsymbol{y} \in \mathbb{R}^m$, and there is a set $Y \subseteq \mathbb{R}^m$ of non-zero measure such that $\mathcal{V}(\mathcal{C}^+(\boldsymbol{y})) < \mathcal{V}(\mathcal{C}^\cap(\boldsymbol{y}))$ for $\boldsymbol{y} \in Y$. Define a new confidence region for the unconstrained problem

$$\mathcal{C}' := \mathcal{C}^+ \cup \mathcal{C}^c, \tag{2.29}$$

where $\mathcal{C}^c = \mathcal{C} \setminus \mathcal{C}^\cap$ denotes the compliment of $\mathcal{C}^\cap$ in $\mathcal{C}$. Then, $\mathcal{C}'$ has the given coverage level, since $\mathcal{C}^+$ provides coverage for $\theta_0 \in \Omega_\mathrm{c}$, whereas $\mathcal{C}^c$ provides coverage for the case $\theta_0 \in \Omega \setminus \Omega_\mathrm{c}$. Furthermore, we have for almost all $\boldsymbol{y}$

$$\begin{aligned} \mathcal{V}(\mathcal{C}'(\boldsymbol{y})) &= \mathcal{V}(\mathcal{C}^+(\boldsymbol{y})) + \mathcal{V}(\mathcal{C}^c(\boldsymbol{y})) \\ &\leq \mathcal{V}(\mathcal{C}^\cap(\boldsymbol{y})) + \mathcal{V}(\mathcal{C}^c(\boldsymbol{y})) \\ &= \mathcal{V}(\mathcal{C}(\boldsymbol{y})). \end{aligned} \tag{2.30}$$

Finally, strict inequality holds in Eq. (2.30) for all $\boldsymbol{y} \in Y$ due to the assumption on $\mathcal{C}^+$. However, this would imply $\mathcal{C}$ not being admissible in contradiction to the assumptions of the Lemma. □

A similar procedure of computing optimal point estimators for constrained problems by modifying an optimal estimator for the unconstrained problem was introduced for the maximum likelihood estimator in [SGS12].

One criticism raised against the use of the truncated confidence regions is the possibility that they may yield empty realizations and, hence, are considered "unphysical" [FC98]. However, according to the standard definition in Section 2.1.1, a procedure that reports 95% confidence regions is allowed to give any result 5% of the time.

Furthermore, a different strategy often adopted for point estimator is to use an unconstrained parametrization for the constrained parameter space. A typical example is a coin toss model with bias $p \in [0, 1]$. Instead of p, the problem can also be parametrized in terms of of log-odds $\log \frac{p}{1-p}$, which can take any value in $(-\infty, \infty)$.

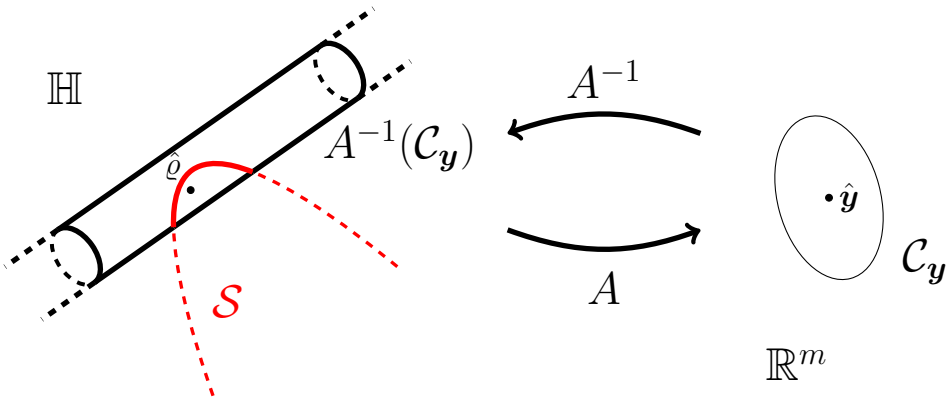


Figure 2.2.: Geometric construction of a confidence region for ϱ_0. Quantum states are mapped by a measurement matrix A to the respective quantum expectation values $\boldsymbol{y}$. Conversely, the pre-image of a confidence region $\mathcal{C}_{\boldsymbol{y}}$ under A gives rise to a confidence region for ϱ_0. These may be unbounded if the measurements are not tomographically complete – a drawback that can be cured by taking into account the physical constraints on quantum states, i.e. positive semi-definiteness.

Similar, one could use the following parametrization for quantum states guaranteed to give a positive semidefinite, Hermitian matrix with trace 1

$$\rho(X) = \frac{XX^\dagger}{\operatorname{tr} XX^\dagger} \tag{2.31}$$

with $X \in \mathbb{C}^{d\times d}$. Although this parametrization can certainly be advantageous for point estimation, it is unlikely to be helpful for uncertainty quantification: While X and $\rho(X)$ carry equivalent information, the size of a region measured in "X-space" is hardly related to the size of a region in the physical state space unless one chooses highly unnatural volume measures. This is necessarily so, as any map from an unbounded space onto the compact quantum state space must grossly distort the geometry. So, having obtained a "small confidence region" in parameter space does not imply that the state has been well-estimated w.r.t. any physically relevant metric.

2.4.2. Confidence regions from linear inversion

A particularly simple approach to QSE is the method of *linear inversion*, which we are going to review now: First, assume that the *true* but unknown quantum state is represented by a $d \times d$ density matrix $\varrho_0 \in \mathcal{S} \subseteq \mathbb{H}$. The data is obtained from measurements of $m \geq d^2 - 1$ tomographically-complete measurement projectors

$E_1, \ldots, E_m$. By $y_k = \operatorname{tr}(E_k \varrho_0)$, $k = 1, \ldots, m$ we denote the (quantum) expectation values of E_k for the true state ϱ_0. Since these relations are linear, we can rewrite them as $\boldsymbol{y} = A\varrho_0$, where A is the measurement (or design) matrix independent of ϱ_0. The (pseudo)-inverse of the above relation is given by

$$\varrho_0 = \left(A^T A\right)^{-1} A^T \boldsymbol{y} \tag{2.32}$$

and simplifies to $\varrho_0 = A^{-1}\boldsymbol{y}$ if $m = d^2 - 1$.

Of course, in an experiment, the expectation values $\boldsymbol{y}$ are unknown and can only be approximated by some estimate $\hat{\boldsymbol{y}}$ based on the observed data. The linear inversion estimate for the quantum state $\hat{\varrho}$ is then given by Eq. (2.32) with the probabilities $\boldsymbol{y}$ replaced by the empirical frequencies $\hat{\boldsymbol{y}}$. However, due to statistical fluctuations the estimated state $\hat{\varrho}$ is not necessarily positive semidefinite [Kni+15], which led to the development of estimators enforcing the physical constraints such as the maximum likelihood estimator [Hra+04]. Although the linear inversion and maximum likelihood estimator solve two distinct problems – namely the unconstrained and constrained one, respectively – in certain cases the two are related. More precisely, if the outcomes approximately follow a Gaussian distribution, a fast projection algorithm computes the maximum likelihood estimate from the linear inversion estimate directly [SGS12].

Here, we take a similar approach. First, the simple geometric interpretation of the linear inversion estimator (see Fig. 2.2) allows us to map confidence regions for the expectation values to confidence regions for the state without taking into account the positivity constraint: If $\mathcal{C}_{\boldsymbol{y}}$ is a confidence region for $\boldsymbol{y}$ with confidence level $1 - \alpha$, then so is its pre-image under the measurement map

$$\mathcal{C}_{\varrho_0} := A^{-1}(\mathcal{C}_{\boldsymbol{y}}) \tag{2.33}$$

for ϱ_0. Second, the truncation $\mathcal{C}_{\varrho_0}^{\cap} := \mathcal{C}_{\varrho_0} \cap \mathcal{S}$ yields an improved confidence region for the problem with quantum constraints taken into account. As shown in Lemma 2.11, this approach yields admissible confidence region provided the original region was admissible.

The same construction can also be carried out for tomographically incomplete measurements, i.e. for $m < d^2$: Since the measurement matrix A is non-invertible in this case, the estimate for the state $\hat{\varrho}$ satisfying $A\hat{\varrho} = \hat{\boldsymbol{y}}$ is not uniquely defined. However, under additional structural assumptions, one can single out a unique estimate [Gro+10; Fla+12]. The singularity of the measurement map A also reflects in the confidence region defined by Eq. (2.33). Even if $\mathcal{C}_{\boldsymbol{y}}$ is a bounded region, the confidence region for the state $\mathcal{C}_{\varrho_0}$ extends to infinity in the directions "unobserved by A". In both cases, the tomographically complete and incomplete, we can use the intersection with the psd states $\mathcal{S}$ to reduce the the region's size while not sacrificing coverage. This improvement is especially far-reaching in the latter case, where it

turns an unbounded region to a bounded one just by taking into account the physical constraints.

Of course, the question is whether we can somehow characterize the truncated confidence region $\mathcal{C}_{\varrho_0}^{\cap} := A^{-1}(\mathcal{C}_{\boldsymbol{y}}) \cap \mathcal{S}$ computationally efficiently. As already mentioned in Section 2.3, we are going to make the simplifying assumption that the measured frequencies are approximately Gaussian distributed. Furthermore, we are going to focus on a class of confidence regions that are efficiently characterizable in the unconstrained setting, namely Gaussian confidence ellipsoids or, more precisely, ellipsoidal balls of the form

$$\mathcal{C}_{\boldsymbol{y}} = \left\{ \boldsymbol{y} \in \mathbb{R}^m \colon (\boldsymbol{y} - \hat{\boldsymbol{y}})^T B \, (\boldsymbol{y} - \hat{\boldsymbol{y}}) \leq 1 \right\} \tag{2.34}$$

centered at the the empirical frequencies $\hat{\boldsymbol{y}}$. The $m \times m$, symmetric, positive semidefinite matrix B completely specifies the ellipsoidal shape of this confidence region. These are the natural generalizations of the well-known 2σ confidence intervals to multivariate Gaussian distributions.

However, even in the unconstrained setting, the ellipsoidal construction (2.34) is known to be admissible only for $m = \{1, 2\}$ [Jos69], while it is not admissible for $m \geq 3$ [Jos67] due to Stein's phenomenon discussed in Section 2.1.1: By shifting the center of the ellipsoid from the empirical mean $\hat{\boldsymbol{y}}$ to the Stein estimator (2.9), one can improve the coverage while keeping the volume constant [Jos67]. Smaller confidence ellipsoids with the same coverage can be obtained by shifting the center slightly [TB+97; HC82] or even by using more complicated shapes [Shi89; BCG95]. Nevertheless, none of these constructions is known to be optimal and, to the best of the author's knowledge, no optimal confidence region for multivariate Gaussians in dimensions $m \geq 3$ is known. But since our discussion is focused on the question how the physical psd constraints can be used to improve confidence regions, we are still going to use the ellipsoids (2.34) as a tractable example: As we will prove later, it is impossible to characterize the truncated ellipsoids efficiently although they are fully described by only few parameters, namely $\hat{\boldsymbol{y}}$ and B.

In the remainder of this section, we are going to discuss a useful parametrization of the ellipsoids $\mathcal{C}_{\varrho_0} = A^{-1}(\mathcal{C}_{\boldsymbol{y}})$ with $\mathcal{C}_{\boldsymbol{y}}$ given by Eq. (2.34). To this end we use the fact that any $d \times d$ Hermitian matrix can be expanded in a basis formed by the identity $\mathbb{1}$ and $d^2 - 1$ traceless Hermitian matrices σ_i, $i = 1, \ldots, d^2 - 1$, normalized according to $\mathrm{Tr}(\sigma_i \sigma_j) = 2\delta_{ij}$. With the symbols σ_i we associate here the most common choice of the basis elements [Kim03] – explicitly provided in Appendix A.1. Any other basis $\sigma_i' = \sum_j O_{ji} \sigma_j$, which can be obtained from σ_i by a $d^2 - 1$ dimensional, orthogonal matrix O, is of course equally valid. For $d = 2$ the choice stated in A.1 is simply the Bloch basis of Pauli matrices: $\sigma_1 = \sigma_x$, $\sigma_2 = \sigma_y$ and $\sigma_3 = \sigma_z$. In higher dimensions

the matrices σ_i maintain the Bloch basis structure:s Let

$$i_d = \frac{d(d-1)}{2}, \tag{2.35}$$

then the definition of σ_i mimics σ_x for $1 \leq i \leq i_d$, σ_y for $i_d + 1 \leq i \leq 2i_d$ and σ_z for $2i_d + 1 \leq i \leq d^2 - 1$. Therefore, we are going to refer to the σ_i as *(generalized) Bloch matrices* and the corresponding parametrization of Hermitian matrices as the *(generalized) Bloch representation*. The following theorem provides a useful parameterization of pre-images under the design matrix A of ellipsoids.

Theorem 2.12. *For the tomographically complete case $m \geq d^2 - 1$, the pre-image under the design matrix of any ellipsoid of the form* (2.34) *can be written as*

$$\mathcal{C}_{\varrho_0} = A^{-1}\left(\mathcal{C}_{\boldsymbol{y}}\right) = \left\{ \hat{\varrho} + \sum_i R_i u_i \sigma_i' \colon \boldsymbol{u}^T \boldsymbol{u} \leq 1 \right\}. \tag{2.36}$$

Here, $\hat{\varrho}$ is the linear inversion estimator corresponding to Eq. (2.32)*, that is a Hermitian matrix with $\operatorname{tr} \hat{\varrho} = 1$. The $R_i > 0$ ($i = 1, \ldots, d^2 - 1$) are the ellipsoid's radii in the directions given by $\sigma_i' = \sum_j O_{ji}\sigma_j$ and the orthogonal matrix $O \in \mathcal{SO}\left(d^2 - 1\right)$ furnishes any orientation of the semi-major axes of the ellipsoid.*

Proof. Note that whenever the sum has no limits specified (like in Eq. (2.36)), by default it extends from 1 to $d^2 - 1$. Let us parameterize both $\varrho \in \mathcal{C}_{\varrho_0}$ and $\hat{\varrho}$ in the Bloch representation with coordinates w_i and $\hat{w}_i$, respectively:

$$\varrho = \frac{1}{d}\mathbb{1} + \sum_i w_i \sigma_i, \qquad \hat{\varrho} = \frac{1}{d}\mathbb{1} + \sum_i \hat{w}_i \sigma_i. \tag{2.37}$$

Since $\boldsymbol{y} = \operatorname{Tr}\left(\boldsymbol{E}\varrho\right)$ and $\hat{\boldsymbol{y}} = \operatorname{Tr}\left(\boldsymbol{E}\hat{\varrho}\right)$ we find

$$\boldsymbol{y} - \hat{\boldsymbol{y}} = Q\left(\boldsymbol{w} - \hat{\boldsymbol{w}}\right), \tag{2.38}$$

where Q is a $m \times (d^2 - 1)$ matrix with elements $Q_{ki} = \operatorname{Tr}\left(E_k \sigma_i\right)$. In other words, the Bloch coordinates satisfy the same ellipsoid equation (2.34) as the measurement outcomes with B substituted by the $d^2 - 1$ dimensional square matrix $B' = Q^T B Q$. Since B is symmetric and positive definite, the same holds for B'. Hence, B' can be diagonalized to the form $B' = ODO^T$, where $O \in \mathcal{SO}(d^2 - 1)$ and $D = \operatorname{diag}(R_1^{-2}, \ldots, R_{d^2-1}^{-2})$ is a diagonal matrix with positive entries. If we rescale $\boldsymbol{w} - \hat{\boldsymbol{w}} = OD^{-1/2}\boldsymbol{u}$, then $\boldsymbol{u}^T\boldsymbol{u} \leq 1$ and

$$\varrho - \hat{\varrho} = \sum_j \left(\sum_i O_{ji} R_i u_i \right) \sigma_j. \tag{2.39}$$

In the last step of the proof we simply change the orientation of the basis to $\sigma_i' = \sum_j O_{ji}\sigma_j$. □

2.4.3. Computational intractability of truncated ellipsoids

Guided by the discussion from the previous section we now study the confidence region for the linear inversion QST defined as

$$\mathcal{C}_{\varrho_0}^{\cap} := \mathcal{C}_{\varrho_0} \cap \mathcal{S} = A^{-1}(\mathcal{C}_{\boldsymbol{y}}) \cap \mathcal{S}, \tag{2.40}$$

where $\mathcal{C}_{\varrho_0}$ is given by the ellipsoid (2.36) for the tomographically complete case $m = d^2 - 1$. In this section, we are going to show that in contrast to the full ellipsoid $\mathcal{C}_{\varrho_0}$, the truncated ellipsoid $\mathcal{C}_{\varrho_0}^{\cap}$ cannot be characterized computationally efficiently. This shows, for example, in the fact that there is no efficient algorithm to answer the following question: How much does taking into account the physical constraints reduce the size of the confidence region on a particular set of observed data? Note that we will not be concerned with properties of the region estimator but with a single instance corresponding to a fixed set of data. By abuse of notation, we are going to refer to these instances as $\mathcal{C}_{\varrho_0}$ and $\mathcal{C}_{\varrho_0}^{\cap}$ as well.

More precisely, we are concerned with the question if a fixed ellipsoid $\mathcal{C}_{\varrho_0}$ changes due to the constraints in Eq. (2.40) or if $\mathcal{C}_{\varrho_0}$ is fully contained in the set of psd states. For the precise formulation, we use the representation of ellipsoids from Thm. 2.12.

Problem 2.13. *Given the center $\hat{\varrho}$, radii R_i, and a basis σ_i' for $\mathbb{H}$. Is there a $\boldsymbol{u} \in \mathbb{R}^{d^2-1}$ with $\boldsymbol{u}^T\boldsymbol{u} \leq 1$ such that*

$$\hat{\varrho} + \sum_i R_i u_i \sigma_i' \in \mathbb{H} \setminus \mathcal{S}? \tag{2.41}$$

The main result of this section is the following statement on the computational complexity of the aforementioned problem.

Theorem 2.14. *Problem 2.13 is* **NP***-complete.*

As a consequence of Theorem 2.14, the problem of "characterizing" the truncated confidence ellipsoids $\mathcal{C}_{\varrho_0}^{\cap} := A^{-1}(\mathcal{C}_{\boldsymbol{y}}) \cap \mathcal{S}$ defined in Sec. 2.4.2 computationally is hard in general. By characterizing we mean computing any property of $\mathcal{C}_{\varrho_0}^{\cap}$ that is sensitive to whether the truncation influences the original ellipsoid or not, e.g. computing the volume of the truncated ellipsoid or its distance to boundary of the quantum state space with high enough precision. Note, however, that there are also properties such as the diameter that might be unaffected by the truncation in certain special cases and, hence, their computational complexity cannot be classified using Theorem 2.14. Furthermore, the more general problem of characterizing truncated confidence regions in general (without the Gaussian approximation) is hard as well since it subsumes Problem 2.13.

Another consequence of the theorem concerns confidence regions for the constrained problem, which output "good regions" for the unconstrained problem when

the constraints are not active: More precisely, it is extremely natural to use likelihood ratio-based ellipsoidal confidence regions for unconstrained Gaussian data although they cannot be optimal due to Stein's phenomenon. So it is natural to require any quantum region estimator to behave this way in the particular case that the likelihood function is concentrated well away from the boundary of state space. What Theorem 2.14 shows is that any region estimator subject to this criterion must necessarily solve NP-hard problems.

2.4.4. Proof of Theorem 2.14

The remainder of this section is dedicated to the proof of Theorem 2.14 and to discuss a tractable special case. First, note Problem 2.13 is in **NP** as any $\boldsymbol{u}$ fulfilling Eq. (2.41) serves as a certificate. The rest of the proof of Theorem 2.14 is inspired by a similar result due to Ben-Tal and Nemirovski [BN98] in robust optimization theory, who showed that the following problem is **NP**-complete.

Problem 2.15. *Given $k \in \mathbb{N}$ and k $d \times d$ symmetric matrices $A_1, \ldots, A_k$, check whether there is a $\boldsymbol{u} \in \mathbb{R}^k$ with $\boldsymbol{u}^T \boldsymbol{u} \leq 1$ such that $\sum_{i=1}^k u_i A_i > \mathbb{1}_d$.*

Although the two problems are strongly related, the intractability result of Problem 2.15 cannot be applied directly to Problem 2.13 due to the following crucial difference: The proof of NP-hardness of Problem 2.15 constructs a reduction of the number partition problem 2.10 to the special case of Problem 2.15 with $k = \frac{d(d-1)}{2} + 1$ and real symmetric matrices A_i, which are not necessarily pairwise orthogonal to each other [BN98, Sec. 3.4.1]. However, in Prob. 2.13, the σ_i' $(i = 1, \ldots, d^2 - 1)$ form an orthogonal basis of the space of complex Hermitian, traceless matrices. Hence, we need to adapt the original proof strategy to deal with the restrictions imposed by our QSE related problem.

For the proof of Theorem 2.14, we show that it is already hard to decide Problem 2.13 for the special case of ellipsoids that have their semi-major axes aligned with the generalized Bloch basis In other words, we assume $\sigma_i' = \sigma_i$. We consider the same radius R_1 for all directions generalizing the x-direction to higher dimensions and the distinct radius R_2 for the remaining directions:

$$\begin{aligned} R_i &= R_1 \quad i = 1, \ldots, i_d \\ R_i &= R_2 \quad i = i_d + 1, \ldots, d^2 - 1. \end{aligned} \tag{2.42}$$

Recall the definition of i_d Eq. (2.35). To prove the hardness of Problem 2.13, we use a reduction from the number partition problem 2.10. The main technical difficulty is to identify the values of R_1 and R_2 as well as $\hat{\rho}$ depending on an instance of the balanced sum problem $\boldsymbol{a}$ such that the corresponding ellipsoid $\mathcal{C}$ given by Theorem 2.12 contains an element with negative eigenvalues if and only if $\boldsymbol{a}$ has a balanced sum

partition.

For this purpose, we introduce an explicit representation of pure states in terms of the orthonormal basis $\{|i\rangle\}_i$ from Appendix A.1. Let $|\Psi\rangle$ denote denote any element from the d dimensional Hilbert space of pure states and define the corresponding complex vector $\boldsymbol{\psi}$ in terms of its coordinates

$$\psi_k = \langle k|\Psi\rangle\,, \qquad k = 1, \ldots, d, \tag{2.43}$$

Consequently, $\sqrt{\langle\Psi|\Psi\rangle}$ is the norm of $|\Psi\rangle$, while $\|\boldsymbol{\psi}\|$ denotes the norm of $\boldsymbol{\psi}$. Obviously both norms are equal.

In a first step of the proof we write down the positivity condition for the ellipsoid under investigation: The confidence ellipsoid $\mathcal{C}$ is fully contained in the set of psd states if and only if for all $\varrho \in \mathcal{C}$ and all $|\Psi\rangle$, $\langle\Psi|\varrho|\Psi\rangle \geq 0$. holds. In the parametrization from Theorem 2.12, this condition can be rewritten as

$$\langle\Psi|\hat{\varrho}|\Psi\rangle + R_1 \sum_{i=1}^{i_d} u_i v_i\left(\boldsymbol{\psi}\right) + R_2 \sum_{i=i_d+1}^{d^2-1} u_i v_i\left(\boldsymbol{\psi}\right) \geq 0, \tag{2.44}$$

where we have already restricted our attention to the special case from Eq. (2.42). Furthermore, we have used the shorthand $v_i\left(\boldsymbol{\psi}\right) = \langle\Psi|\sigma_i|\Psi\rangle$, which are the rescaled Bloch coordinates of the density matrix $|\Psi\rangle\langle\Psi|$. Condition (2.44) is independent of the norm of $|\Psi\rangle$. Thus, we can fix $\langle\Psi|\Psi\rangle = d$. Recall that Eq. (2.44) has to hold for all values of $\boldsymbol{u}$ with $\boldsymbol{u}^T\boldsymbol{u} \leq 1$. Since the left hand side assumes its minimal value for

$$u_i = -\frac{v_i\left(\boldsymbol{\psi}\right)}{\sqrt{\sum_j v_i^2\left(\boldsymbol{\psi}\right)}}, \tag{2.45}$$

we find that Eq. (2.44) is equivalent to

$$\langle\Psi|\hat{\varrho}|\Psi\rangle - \sqrt{R_1^2 \sum_{i=1}^{i_d} v_i^2\left(\boldsymbol{\psi}\right) + R_2^2 \sum_{i=i_d+1}^{d^2-1} v_i^2\left(\boldsymbol{\psi}\right)} \geq 0. \tag{2.46}$$

Using the unusual normalization of $|\Psi\rangle$, we find

$$\sum_i v_i^2\left(\boldsymbol{\psi}\right) = 2d\left(d-1\right) =: \mathcal{P}, \tag{2.47}$$

which can be utilized to simplify (2.46)

$$g(\boldsymbol{\psi}) := \langle\Psi|\hat{\varrho}|\Psi\rangle - \sqrt{\mathcal{P}R_2^2 + \left(R_1^2 - R_2^2\right)\sum_{i=1}^{i_d} v_i^2\left(\boldsymbol{\psi}\right)} \geq 0. \tag{2.48}$$

In the following, we restrict our attention to $R_1 > R_2$, so that both term inside the square root are manifestly positive.
In the second step of the proof we show and utilize the following lemma:

Lemma 2.16. *If $\hat{\varrho}$ is a real, symmetric matrix w.r.t. $|i\rangle$, then the minimum of $g(\boldsymbol{\psi})$ is attained by a vector $\boldsymbol{\psi}$ with real coordinates.*

Proof. Note that we can decompose any vector $|\Psi\rangle$ into its real and imaginary part

$$|\Psi\rangle = |\Psi_1\rangle + \mathrm{i}\,|\Psi_2\rangle\,, \tag{2.49}$$

where the $|\Psi_i\rangle$ are defined in terms of real vectors $\boldsymbol{\psi}_i$. Therefore, for $\hat{\varrho}$ being real and symmetric, we find

$$\langle\Psi|\hat{\varrho}|\Psi\rangle = \langle\Psi_1|\hat{\varrho}|\Psi_1\rangle + \langle\Psi_2|\hat{\varrho}|\Psi_2\rangle. \tag{2.50}$$

A similar equality holds with $\hat{\varrho}$ replaced by $\mathbb{1}$ or σ_i for $i = 1, \ldots, i_d$, since the latter matrices are symmetric and real as well. To shorten the notation, we now define two $i_d + 1$ dimensional vectors $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ with components ($\alpha = 1, 2$)

$$\begin{aligned} x_0^\alpha &= \frac{\sqrt{\mathcal{P}}}{d} R_2 \|\boldsymbol{\psi}_\alpha\|^2 \\ x_i^\alpha &= \sqrt{R_1^2 - R_2^2}\; v_i\left(\boldsymbol{\psi}_\alpha\right) \qquad (i = 1, \ldots, i_d). \end{aligned} \tag{2.51}$$

Since $d = \|\boldsymbol{\psi}\|^2 = \|\boldsymbol{\psi}_1\|^2 + \|\boldsymbol{\psi}_2\|^2$, we find

$$\sqrt{\mathcal{P} R_2^2 + (R_1^2 - R_2^2) \sum_{i=1}^{i_d} v_i^2\left(\boldsymbol{\psi}\right)} = \|\boldsymbol{x}^1 + \boldsymbol{x}^2\| \leq \|\boldsymbol{x}^1\| + \|\boldsymbol{x}^2\|, \tag{2.52}$$

where we used triangle inequality in the last step. Therefore we have

$$g(\boldsymbol{\psi}) \geq g(\boldsymbol{\psi}_1) + g(\boldsymbol{\psi}_2) \tag{2.53}$$

Equation (2.53) implies that if $g(\boldsymbol{\psi})$ is non-negative for all real vectors then it is also non-negative for any vector $\boldsymbol{\psi}$. More intuitively, the above result is true because the construction of $g(\boldsymbol{\psi})$ utilizes only the generalized σ_{X} Pauli matrices and the expectation values of such generalized Pauli matrices are depend on the real parts of $\boldsymbol{\psi}^* \otimes \boldsymbol{\psi}$ w.r.t. the fixed basis. The imaginary part of such projectors only show up in expectation values of the generalized σ_{Y} matrices. □

The next step of the proof, which is crucial for encoding an instance number partition problem, is the choice of the ellipsoid's center $\hat{\varrho}$. We choose

$$\hat{\varrho} = \frac{q}{d}\mathbb{1} + \frac{1-q}{a^2}|\boldsymbol{a}\rangle\langle\boldsymbol{a}|, \qquad 0 \leq q \leq 1, \qquad a = \|\boldsymbol{a}\|, \tag{2.54}$$

with $q \in \mathbb{R}$ to be specified below and $|\boldsymbol{a}\rangle = \sum_k a_k |k\rangle$ denoting a state represented by a real, *integral* vector $\boldsymbol{a}$. The latter are exactly the instances of the number partition problem 2.10. Since $\hat{\varrho}$ given by Eq. (2.54) is manifestly real and symmetric, we can restrict our attention to $\boldsymbol{\psi} \in \mathbb{R}^d$ due to Lemma 2.16. We find

$$\langle \Psi | \hat{\varrho} | \Psi \rangle = q + \frac{1-q}{a^2} (\boldsymbol{a} \cdot \boldsymbol{\psi})^2, \tag{2.55}$$

and

$$\sum_{i=1}^{i_d} v_i^2 (\boldsymbol{\psi}) = 4 \sum_{1 \leq j < k \leq d} \psi_j^2 \psi_k^2 \equiv 2d^2 - 2 \sum_{k=1}^{d} \psi_k^4. \tag{2.56}$$

Before we will be ready to take an advantage of the above encoding we need to perform a sequence of tedious algebraic manipulations. In short, the function we work with has an algebraic form $g(\boldsymbol{\psi}) = \kappa - \sqrt{\Delta}$, with both κ and Δ being non-negative. Testing if this function is non-negative is thus equivalent to checking the inequality $\kappa^2 - \Delta \geq 0$. If we divide this inequality by $2(R_1^2 - R_2^2)$ and fix $q = q_+$ or $q = q_-$ with

$$q_\pm = \frac{1}{2} \left(1 \pm \sqrt{1 - 8d \left(R_1^2 - R_2^2 \right) \frac{a^2}{1+a^2}} \right). \tag{2.57}$$

we can rearrange it to the convenient form

$$f (\boldsymbol{\psi}) - C_2 (\boldsymbol{a} \cdot \boldsymbol{\psi})^4 \leq C_1, \tag{2.58}$$

where:

$$f (\boldsymbol{\psi}) = 2d^2 - \sum_{k=1}^{d} \psi_k^4 - 2d \frac{(\boldsymbol{a} \cdot \boldsymbol{\psi})^2}{1+a^2}, \tag{2.59}$$

$$C_1 = d^2 + \frac{1}{R_1^2 - R_2^2} \left[\frac{q_\pm^2}{2} - d (d-1) R_2^2 \right], \tag{2.60}$$

$$C_2 = \frac{q_\mp^2}{2a^4 \left(R_1^2 - R_2^2 \right)} > 0 \tag{2.61}$$

Both solutions (2.57) assure that (2.58) is free from additional terms proportional to $(\boldsymbol{a} \cdot \boldsymbol{\psi})^2$, except those already included in f.

Hence, the original problem of deciding whether the ellipsoid $\mathcal{E}$ centered at $\hat{\varrho}$ and with radii (2.42) is contained in the psd states can be rephrased as deciding whether the maximum of the left hand side of Eq. (2.58) is smaller or equal to some constant:

$$\mathcal{E} \subseteq \mathcal{S} \iff \max_{\boldsymbol{\psi} \in \mathbb{S}_d^{d-1}} \left[f (\boldsymbol{\psi}) - C_2 (\boldsymbol{a} \cdot \boldsymbol{\psi})^4 \right] \leq C_1. \tag{2.62}$$

Here, $\mathbb{S}_{\zeta}^{d-1}$ denotes a $(d-1)$-dimensional sphere in $\mathbb{R}^d$ with radius $\sqrt{\zeta}$, i.e.

$$\boldsymbol{\psi} \in \mathbb{S}_d^{d-1} \iff \boldsymbol{\psi} \in \mathbb{R}^d \ \wedge \ \|\boldsymbol{\psi}\|^2 = d. \tag{2.63}$$

The relation of Problem 2.13 to the number partition problem is derived in the following Lemma.

Lemma 2.17. *If the instance $\boldsymbol{a}$ of Problem 2.10 allows for a partition, then*

$$\max_{\boldsymbol{\psi}\in\mathbb{S}_d^{d-1}} \left[f(\boldsymbol{\psi}) - C_2(\boldsymbol{a}\cdot\boldsymbol{\psi})^4 \right] = 2d^2 - d. \tag{2.64}$$

On the other hand, if there is no such partition, we have

$$\max_{\boldsymbol{\psi}\in\mathbb{S}_d^{d-1}} \left[f(\boldsymbol{\psi}) - C_2(\boldsymbol{a}\cdot\boldsymbol{\psi})^4 \right] < \max_{\boldsymbol{\psi}\in\mathbb{S}_d^{d-1}} f(\boldsymbol{\psi}) \tag{2.65}$$

$$\leq 2d^2 - d - \frac{2}{p(ad)}. \tag{2.66}$$

where $p(x) = 2x^4$ is a non-negative polynomial.

For the sake of clarity we relegate the proof of the above lemma to the end of this section and discuss its implications now. As a consequence of Lemma 2.17 the choice,

$$C_1 = 2d^2 - d - p(ad)^{-1}, \tag{2.67}$$

implies that an efficient algorithm deciding whether the inequality (2.58) is satisfied or not is also capable of deciding the number partition problem 2.10 efficiently. This proves the claim of Theorem 2.14 that Problem 2.13 is **NP**-hard.

The last step we need to make is to find the parameters R_1 and R_2 leading to the choice (2.67). To this end, we set $R_2 = \epsilon R_1$ with $0 < \epsilon < 1$ and introduce two positive parameters

$$B_1 = p(ad)^{-1}, \qquad B_2 = \frac{da^2}{1+a^2}. \tag{2.68}$$

Note that if $1 \leq j \leq d$ is such that $|a_j| = \min_k |a_k|$, then for $\boldsymbol{\psi}^j$ given by $\psi_k^j = \sqrt{d}\delta_{jk}$ the function $f(\boldsymbol{\psi}^j)$ is equal to

$$f\left(\boldsymbol{\psi}^j\right) = \frac{d^2}{1+a^2}\left(1 + a^2 - 2a_j^2\right). \tag{2.69}$$

Since $a^2 - 2a_j^2 \geq (d-2)a_j^2$ the quantity $f(\boldsymbol{\psi}^j)$ is non-negative, so is the right hand side of Eq. (2.65). From (2.66) we can find the bound

$$B_1 \leq d^2 - d/2. \tag{2.70}$$

Furthermore, $B_2 \leq d$.

Rearranging Eq. (2.60), taking the square root and substituting (2.67) we can see that R_1 is implicitly defined by the relation

$$\sqrt{2}\sqrt{\left(d^2 - d - B_1\right)\left(1-\epsilon^2\right) + d\left(d-1\right)\epsilon^2}R_1 = q_{\pm}. \tag{2.71}$$

If the left hand side of (2.71) happens to be bigger than $1/2$, we need to take the q_+ solution on the right hand side (and q_- in the opposite case). In order for the square roots in Eq. (2.71) to be real-valued, we need to assume

$$\left(d^2 - d - B_1\right)\left(1-\epsilon^2\right) + d\left(d-1\right)\epsilon^2 \geq 0. \tag{2.72}$$

and

$$1 - 8R_1^2\left(1-\epsilon^2\right)B_2 \geq 0, \tag{2.73}$$

The latter condition assures that $q_{\pm}$ are real while the former condition, as it does not depend on R_1, can be immediately solved for ϵ:

$$\epsilon^2 \geq 1 - \frac{d\left(d-1\right)}{B_1}. \tag{2.74}$$

However, Eq. (2.74) does not yield a universal bound for acceptable values of ϵ since B_1 depends on the particular instance $\boldsymbol{a}$. To obtain a lower bound independent of $\boldsymbol{a}$, we use Eq. (2.70), obtaining:

$$\epsilon^2 \geq \frac{1}{2d-1}. \tag{2.75}$$

Since both sides of (2.71) are non-negative, we can take the square of this relation and turn it it into a quadratic equation for R_1. Surprisingly, this equation has a trivial solution $R_1 = 0$ (only relevant while dealing with q_-) and a single non-trivial solution which can be simplified to the form:

$$R_1 = \frac{1}{\sqrt{2}}\frac{\sqrt{d\left(d-1\right) - B_1\left(1-\epsilon^2\right)}}{d\left(d-1\right) - \left(B_1 - B_2\right)\left(1-\epsilon^2\right)}, \tag{2.76}$$

The condition (2.73) becomes trivially satisfied, while the left hand side of Eq. (2.71) is greater than $1/2$ (relevant for q_+) for

$$\epsilon^2 \geq 1 - \frac{d\left(d-1\right)}{\left(B_1 + B_2\right)}. \tag{2.77}$$

In the opposite case the inequality is reversed. When (2.77) occurs, we find that

$$q_+ = \frac{d\left(d-1\right) - B_1\left(1-\epsilon^2\right)}{d\left(d-1\right) - \left(B_1 - B_2\right)\left(1-\epsilon^2\right)}, \tag{2.78}$$

$$q_- = \frac{B_2\left(1-\epsilon^2\right)}{d\left(d-1\right) - \left(B_1 - B_2\right)\left(1-\epsilon^2\right)}, \tag{2.79}$$

while in the opposite case the parameters q_+ and q_- swap. These interrelations between the parameters imply that regardless of the validity of (2.77), the solution (2.76) uniquely determines q initially introduced in (2.54) as given by the formula (2.78). This parameter is manifestly smaller than 1 and due to (2.74) it is also non-negative. With the given choice of parameters (2.76) and (2.77) as well as q specified above, we complete the reduction from the number partition problem. To finalize the proof of Theorem 2.14, we now state the proof of Lemma 2.17.

Proof of Lemma 2.17. The first part of the proof – Eq. (2.64) – follows from a simple calculation utilizing the partition vector $\boldsymbol{\psi}$ defined in (2.23). Note that as $\boldsymbol{a} \cdot \boldsymbol{\psi} = 0$, we immediately obtain the first equality in (2.64), which since C_2 is non-negative turns into inequality in (2.65).

To prove (2.66), we define the set of all possible (2^d in total) partition vectors

$$\mathcal{Z} := \left\{ \boldsymbol{z} \in \mathbb{R}^d \colon \forall i \, z_i = \pm 1 \right\} \tag{2.80}$$

and (for an arbitrary $0 < \lambda < 1$) the set of vectors that are "close" to some element from $\mathcal{Z}$

$$\mathcal{B} := \left\{ \boldsymbol{\psi} \in \mathbb{R}^d \colon \min_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{\psi} - \boldsymbol{z}\| \leq \frac{\lambda}{a} \right\}. \tag{2.81}$$

Because $a \geq 1$, the set $\mathcal{B}$ can be thought of as a disjoint union of 2^d balls centered around the elements of $\mathcal{Z}$. For further convenience we denote $\tilde{\boldsymbol{z}} = \operatorname{argmin}_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{\psi} - \boldsymbol{z}\|$, and $\boldsymbol{\delta} := \boldsymbol{\psi} - \tilde{\boldsymbol{z}}$. By construction $\tilde{z}_k = \operatorname{sign} \psi_k$ so that for all $k = 1, \ldots, d$

$$\tilde{z}_k \delta_k = \tilde{z}_k \psi_k - \tilde{z}_k^2 = |\psi_k| - 1 \geq -1. \tag{2.82}$$

Since $\|\boldsymbol{\psi}\|^2 = d$ we find that

$$2\tilde{\boldsymbol{z}} \cdot \boldsymbol{\delta} = -\|\boldsymbol{\delta}\|^2. \tag{2.83}$$

Using all the above, the fact that $\tilde{z}_k^2 = 1$ and $\tilde{z}_k^3 = \tilde{z}_k$, and the Jensen inequality we can further estimate

$$-\sum_{k=1}^{d} \psi_k^4 \leq -d - \sum_{k=1}^{d} \delta_k^4 \leq -d - \frac{\|\boldsymbol{\delta}\|^4}{d}. \tag{2.84}$$

As $\boldsymbol{a}$ does not allow for partition and both, $\tilde{\boldsymbol{z}}$ and $\boldsymbol{a}$ are integral, we must necessarily have $|\boldsymbol{a} \cdot \tilde{\boldsymbol{z}}| \geq 1$. Thus

$$1 \leq |\boldsymbol{a} \cdot \tilde{\boldsymbol{z}}| = |\boldsymbol{a} \cdot (\boldsymbol{\psi} - \boldsymbol{\delta})| \leq |\boldsymbol{a} \cdot \boldsymbol{\psi}| + |\boldsymbol{a} \cdot \boldsymbol{\delta}| \leq |\boldsymbol{a} \cdot \boldsymbol{\psi}| + a\|\boldsymbol{\delta}\|, \tag{2.85}$$

so that

$$-|\boldsymbol{a} \cdot \boldsymbol{\psi}| \leq \min\{0, a\|\boldsymbol{\delta}\| - 1\}, \tag{2.86}$$

Taking all the above results together with $|\boldsymbol{a}\cdot\boldsymbol{\psi}| \leq a\|\boldsymbol{\psi}\| = a\sqrt{d}$ we obtain

$$f(\boldsymbol{\psi}) \leq 2d^2 - d - \frac{\|\boldsymbol{\delta}\|^4}{d} + 2d^{3/2}a\frac{\min\{0, a\|\boldsymbol{\delta}\| - 1\}}{1+a^2}. \tag{2.87}$$

We will now study two cases. For $\psi \in \mathcal{B}$, we have $0 \leq \|\boldsymbol{\delta}\| \leq \lambda/a$, so that

$$f(\boldsymbol{\psi}) \leq 2d^2 - d - 2d^{3/2}a\frac{1-\lambda}{1+a^2}, \tag{2.88}$$

while for the opposite case ($\psi \notin \mathcal{B}$), when $\|\boldsymbol{\delta}\| > \lambda/a$, one finds

$$f(\boldsymbol{\psi}) \leq 2d^2 - d - \frac{\lambda^4}{da^4}. \tag{2.89}$$

Therefore, we have for any $\boldsymbol{\psi} \in \mathbb{R}^d$ with $\|\boldsymbol{\psi}\|^2 = d$

$$f(\boldsymbol{\psi}) \leq 2d^2 - d - \min\left\{2d^{3/2}a\frac{1-\lambda}{1+a^2}, \frac{\lambda^4}{da^4}\right\}, \tag{2.90}$$

so that by setting $\lambda = d^{-3/4}$ we obtain the desired result with $p(ad) = 2(ad)^4$. □

We conclude this section by investigating a tractable special case of Problem 2.13. Consider $R_i = R$ for all $i = 1, \ldots, d^2 - 1$ that is all radii are equal and the ellipsoid is a ball. With no loss of generality, we can assume $\sigma_i' = \sigma_i$. The following Lemma provides an easily checkable, necessary, and sufficient condition to decide Problem 2.13 for this special case.

Lemma 2.18. *Let $\mathcal{C}$ denote a ball parameterized according to Theorem 2.12 with with radii $R_i = R$ and midpoint $\hat{\varrho}$. $\mathcal{C}$ is fully contained in the set of psd density matrices if and only if*

$$R \leq \sqrt{\frac{d}{2\,(d-1)}}\ \text{mineig}\,\hat{\varrho}, \tag{2.91}$$

where mineig $\hat{\varrho}$ *denotes the smallest eigenvalue of* $\hat{\varrho}$.

Proof. To check whether a sphere with radius R centered at $\hat{\varrho}$ is contained in the set of psd states, specialize Eq. (2.46) to the special case $R_1 = R_2$:

$$\langle\Psi|\hat{\varrho}|\Psi\rangle - R\sqrt{\sum_i v_i^2\,(\boldsymbol{\psi})} \geq 0. \tag{2.92}$$

Since for any pure state $|\Psi\rangle$ the identity

$$\sum_i v_i^2\,(\boldsymbol{\psi}) = \frac{2\,(d-1)}{d}, \tag{2.93}$$

holds (Bloch vectors of pure states live on the hypersphere), the inequality in question becomes

$$\langle\Psi|\hat{\varrho}|\Psi\rangle - R\sqrt{\frac{2\,(d-1)}{d}} \geq 0. \tag{2.94}$$

Simple minimization with respect to $|\Psi\rangle$ concludes the proof. □

The statement is a straightforward but interesting extension of the known result that the largest ball centered at the completely mixed state and fully contained in the set of psd density matrices has radius $R_{\max} = \sqrt{\frac{1}{2d(d-1)}}$. Intuitively, when the center of the ball is moved away from the completely mixed state, the allowed radii become smaller. This correction happens to be quantified by the smallest eigenvalue of the new center. In conclusions, spherical ellipsoids do not constitute hard instances of Problem 2.13 provided that the minimal eigenvalue of $\hat{\varrho}$ can be computed efficiently with high enough accuracy. However, the slight modification with two distinct radii considered above

2.5. Hardness results for credible regions

In this section we are going to present the second major result of [SRG+17] concerned with Bayesian credible regions in QSE. Bayesian techniques in the context of QSE have been introduced by different authors [Jon91; Sla95; Der+97; SBC01; Buž+98]. The main advantage the Bayesian approach to QSE has over the more established frequentist's methods is the ability to include prior knowlege naturally into the inference procedure. Furthermore, the Bayesian framework is conceptually simpler and it provides notion of error region with a more natural interpretation as discussed in Section 2.5. However, analytical solutions for Bayesian inference problems only exist in special cases when we consider conjugate priors for the likelihood functions. In [AS09], the authors use the self-conjugate Gaussian priors to provide an approximative estimate of the state and the corresponding credible region. However, in order to deal with the psd constraints, the authors had to resort to an uncontrolled approximation. In this work we use the same model and show that the *exact* inference problem under quantum shape constraints cannot be solved efficiently. A different approach, which was pioneered in [HH12; Fer14b], is to use Monte Carlo algorithms to perform approximate inference. Such sampling algorithms also allow the computation of highest posterior density ellipsoids, which are near-optimal credible region for the posteriors relevant to QSE, i.e. Pauli measurements on multiple qubits [Fer14a]. For more details on the practical application of Bayesian inference to QSE, we refer the reader to [GFF17].

2.5.1. MVCR for Gaussian distributions

As a first step towards obtaining good credible regions for the QSE model we ignore the positivity constraints in this section. The Gaussian approximation of the measurement statistics in Eq. (2.27) suggests that we use a Gaussian prior for ϱ: On the one hand, this yields a Gaussian posterior as well and the parameters can be computed analytically by means of *linear Kalman filter update equations* similar to Eq. (2.14) [AS09, Sec. 2.4]. On the other hand, the MVCR for Gaussian distributions are simply ellipsoids, and therefore, can be characterized by a few parameters. Below, we show how those parameters can be computed efficiently.

Since credible regions are purely defined in terms of the posterior, we can ignore the details how the posterior arose. Hence, we assume that the posterior distribution of ϱ under consideration is a Gaussian with mean θ and covariance matrix Σ. In other words, we assume the posterior for ϱ has probability density

$$\Pi_{\theta,\Sigma}(\varrho) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\varrho - \theta\|_\Sigma^2\right). \tag{2.95}$$

where

$$\|\varrho - \theta\|_\Sigma := \sqrt{(\varrho-\theta)^T \Sigma^{-1} (\varrho - \theta)} \tag{2.96}$$

is the Mahalanobis distance and $|\Sigma|$ denotes the determinant of Σ. As elaborated in Section 2.1.2, the MVCRs are exactly the highest posterior density sets as defined in Eq. (2.17). Therefore, the MVCR with credibility α for the Gaussian posterior (2.95) is given by

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^N \colon \|\boldsymbol{x} - \boldsymbol{\theta}\|_\Sigma \leq r_\alpha\} =: \mathcal{E}(r_\alpha). \tag{2.97}$$

This is an ellipsoid centered at $\boldsymbol{\theta}$ with radius r_α determined by the saturated credibility condition (2.16):

$$\begin{aligned} \alpha &= (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \int_{\mathcal{E}(r_\alpha)} \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\theta}\|_\Sigma^2\right) \mathrm{d}^N x \\ &= \frac{\gamma\left(\frac{N}{2}, \frac{r_\alpha^2}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} \equiv P\left(\tfrac{N}{2}, \tfrac{r_\alpha^2}{2}\right). \end{aligned} \tag{2.98}$$

By $\gamma(\cdot,\cdot)$ we denote the incomplete Γ-function and $P(\cdot,\cdot)$ is its normalized version. The above condition fixes r_α uniquely since $x \mapsto P(\frac{N}{2}, x)$ is strictly monotonic for any $N > 0$. Hence, determining the MVCR for a multivariate Gaussian posterior with known mean and covariances reduces to computing the radius r_α, which is formalized in the following problem.

Problem 2.19. *For given mean $\boldsymbol{\theta} \in \mathbb{R}^N$, covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$ with $\Sigma \geq 0$, credibility $\alpha \in [0,1]$, and accuracy δ with $\delta^{-1} \in \mathbb{N}$, determine the radius of the MVCR r_α defined in Eq.* (2.98) *with given accuracy.*

Note that this does not constitute a decision problem, but an algebraic computation: We are asking to compute a number y such that

$$y \in (r_\alpha - \delta, r_\alpha + \delta). \tag{2.99}$$

Although there are elaborate algebraic computational models [AB09, Sec. 16], we are going to use the following simpler strategy here: We are going to consider algorithms for problems such as Problem 2.19 that compute a rational number y with binary encoding of its numerator and denominator. Such an algorithm exists for solving Problem 2.19 that runs in polynomial time in the sense of Definition 2.5. With the shorthand notation $x = r_\alpha^2/2$, the algorithm is outlined below:

1. W.l.o.g. we can assume that $\alpha \leq 0.9$ (or some other arbitrary constant). Otherwise, the problem can be restated in terms of $Q(\frac{N}{2}, x) = 1 - P(\frac{N}{2}, x)$, which allows for a similar analysis. The condition $\alpha \leq 0.9$ restricts the search space for x to some finite interval $[0, t_{\max}]$. Note that the upper bound $t_{\max}$ grows at worst polynomially in $\frac{N}{2}$.

2. The above restriction, the finite precision, and the fact that $x \mapsto P(\frac{N}{2}, x)$ is strictly monotonic allow for interpreting the problem of finding x given α as a search in an ordered, finite list of size $M \sim \frac{t_{\max}}{\delta}$.

3. Each entry of this list can be evaluated with exponential precision in polynomial time using a power series expansion of $P(\frac{N}{2}, x)$ (for more details see Lemma 2.27 in 2.5.4).

4. Since finding x in this list only requires $\log M$ evaluations using binary search, the whole problem can be solved in polynomial time.

2.5.2. Bayesian QSE

In order to incorporate the positivity constraints on ϱ imposed by quantum mechanics, we choose a prior distribution that is concentrated on $\mathcal{S}$ and vanishes on its complement. One possible choice are truncated Gaussian priors. These are defined in terms of their density $\Pi^+_{\theta,\Sigma}(\varrho)$ with respect to the flat Hilbert-Schmidt measure $\mathrm{d}\varrho$ on $\mathbb{H}$

$$\Pi^+_{\theta,\Sigma}(\varrho) = C_{\theta,\Sigma}\ \chi(\varrho)\ \Pi_{\theta,\Sigma}(\varrho). \tag{2.100}$$

Here, $\Pi_{\theta,\Sigma}$ is the multivariate Gaussian from Eq. (2.95) with $\theta \in \mathbb{H}$. The other factors in Eq. (2.100) ensure that $\Pi^+_{\theta,\Sigma}$ is a proper probability distribution supported on $\mathcal{S}$: $\chi(\varrho)$ is the indicator function of $\mathcal{S}$ and $C_{\theta,\Sigma}$ is the normalization constant defined by

$$C^{-1}_{\theta,\Sigma} = \int_{\mathcal{S}} \Pi_{\theta,\Sigma}(\varrho)\, \mathrm{d}\varrho. \tag{2.101}$$

Since the numerator in Bayes rule (2.12) is linear in the prior, the posterior distribution updated with Gaussian likelihood function is also of the form (2.100). Furthermore, we can use the same linear Kalman filter update equations for both the standard and the truncated Gaussian distributions. The only additional complication of computing the posterior corresponding to the truncated prior (2.100) is that the normalization factor $C_{\theta,\Sigma}$ needs to be reevaluated after each update step in order to obtain a properly normalized probability distribution. From now on we only consider a fixed posterior distribution and drop the subscripts indicating the mean θ and the covariance matrix Σ if no confusion arises. It is then important to remember that the constant in question is denoted by C, while the credibility region is $\mathcal{C}$.

The problem we try to solve is the following: Given the mean θ, covariance matrix Σ, and credibility α, can we find the MVCR for the truncated Gaussian distribution supported on $\mathcal{S}$? Since the truncated density (2.100) is supported on the psd states and MVCRs are highest-density sets due to (2.17), the MVCR is of the form

$$\mathcal{E}(r_\alpha^+) \cap \mathcal{S} = \{\varrho \in \mathcal{S} \colon \|\varrho - \theta\|_\Sigma \leq r_\alpha^+\}. \tag{2.102}$$

Similar to Eq. (2.98), the radius is determined by the credibility condition

$$\alpha = C \int_{\mathcal{E}(r_\alpha^+)\cap\mathcal{S}} \Pi_{\theta,\Sigma}(\varrho)\mathrm{d}\varrho. \tag{2.103}$$

However, this case involves the normalization constant C from (2.100) and the integral is restricted to the psd states. Also, there is no closed-form analogue to Eq. (2.98) due to the psd constraint.

2.5.3. Computational intractability

Our main result from this section concerns MVCR for Gaussian posteriors that are fully supported on the psd states. We will show that the following problem is computational hard.

Problem 2.20. *For given mean $\theta \in \mathbb{H}$, covariance matrix Σ, credibility $\alpha \in [0,1]$, and accuracy δ with $\delta^{-1} \in \mathbb{N}$, determine the radius of the MVCR r_α^+ defined in Eq.* (2.103) *with given accuracy.*

In other words, there is no efficient algorithm that outputs smallest volume credibility regions for every Gaussian distribution on $\mathbb{H}$ restricted to the positive semidefinite states and every credibility α. Consequently, there cannot be an efficient algorithm to solve the problem of MVCR for QSE, since the latter more general problem contains the instances of Problem 2.20. To prove Problem 2.20, we use a reduction from Problem 2.13, which has already been shown to be **NP**-complete. This reduction runs along the following lines:

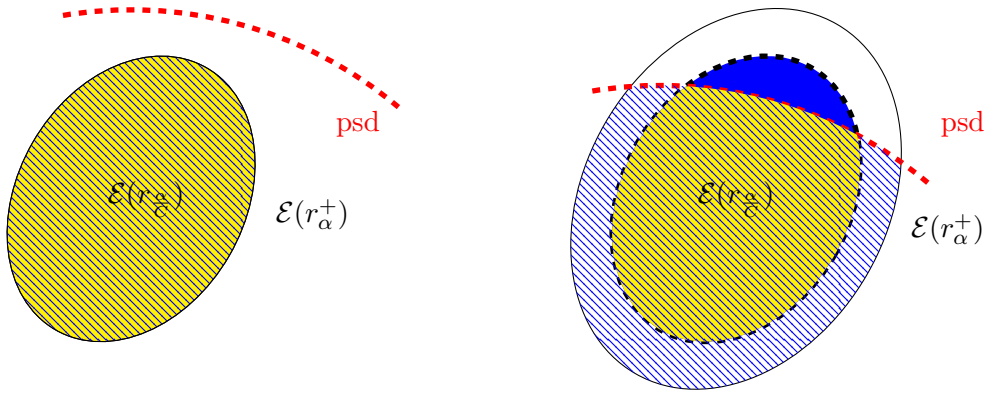


Figure 2.3.: The two possible cases for the credible regions. *Left:* The original ellipsoid $\mathcal{E}(r_{\frac{\alpha}{C}})$ with credibility $\frac{\alpha}{C}$ (yellow) lies completely inside the psd states and is, therefore, equal to the ellipsoid taking into account positivity $\mathcal{E}(r_\alpha^+)$ with credibility α (blue hatched). *Right:* Parts of the original ellipsoid $\mathcal{E}(r_{\frac{\alpha}{C}})$ lie outside the psd states (blue). Hence, the ellipsoid that takes into account positivity $\mathcal{E}(r_\alpha^+)$ has to have a larger radius in order to achieve the sought for credibility.

1. Assume that Prob. 2.20 can be solved efficiently.

2. As we will prove later, every ellipsoid $\mathcal{E}^*$ in $\mathbb{H}$ can be encoded as a minimum volume credible ellipsoid for some Gaussian distribution Π with a suitable choice of θ, Σ, and R:

$$\mathcal{E}^* = \mathcal{E}_{\theta,\Sigma}(R). \tag{2.104}$$

   Note that only θ is uniquely defined. Σ is defined only up to a multiplicative, positive constant, since every rescaling of Σ can be compensated by an appropriate rescaling of R.

3. Using the assumed efficient algorithm for Prob. 2.20, we can compute the normalization constant C of the truncated distribution (2.100) for given θ and Σ with sufficient precision in polynomial time.

4. Based on this, we can compute a credibility α such that $R = r_{\frac{\alpha}{C}}$ and, therefore,

$$\mathcal{E}^* = \mathcal{E}_{\theta,\Sigma}(r_{\frac{\alpha}{C}}). \tag{2.105}$$

5. The crucial observation is that this ellipsoid is contained in the psd states if and only if the corresponding MVCR for the truncated distribution Π^+ fulfills

$$r_\alpha^+ = r_{\frac{\alpha}{C}}. \tag{2.106}$$

See Fig. 2.3 for an illustration. Since we can compute r_α^+ efficiently by assumption, checking Eq. (2.106) allows us to decide Prob. 2.13.

In conclusion, the main result from this section is the following lower bound on the computational complexity of Problem 2.19.

Theorem 2.21. *If Problem 2.20 has a polynomial time algorithm, then we can also decide Problem 2.13 in polynomial time. Therefore, there is no efficient algorithm for Problem 2.20 unless* P = NP*.*

The proof runs along the lines outlined above and can be found in the next section. The main technical problem is that we are dealing with finite-precision arithmetic.

2.5.4. Proof of Theorem 2.21

Let us now construct the polynomial time reduction of Problem 2.13 to Problem 2.19. We will begin with the main observation of this proof, namely Eq. (2.106).

Lemma 2.22. *Let* $\Pi(\varrho)$ *denote a Gaussian distribution on* $\mathbb{H}$ *and* $\Pi^+(\varrho) = C\Pi(\varrho)\chi(\varrho)$ *the corresponding restricted Gaussian with the same mean and covariance matrix, as defined in Eq.* (2.100)*. For any* $\alpha \in [0,1]$*, the credible ellipsoid* $\mathcal{E}(r_{\frac{\alpha}{C}})$ *with credibility* $\frac{\alpha}{C}$ *is contained in the psd if and only if the credible ellipsoid for* Π^+*,* $\mathcal{E}(r_\alpha^+)$*, with credibility* α *has the same radius, that is Eq.* (2.106) *holds.*

Proof. The two cases of $\mathcal{E}(r_{\frac{\alpha}{C}})$ being contained and not being contained in the psd states are illustrated in Fig. 2.3. First, assume that $\mathcal{E}(r_{\frac{\alpha}{C}}) \subseteq \mathcal{S}$, then

$$\frac{\alpha}{C} = \int_{\mathcal{E}(r_{\frac{\alpha}{C}})} \Pi(\varrho)\,\mathrm{d}\varrho. \implies \alpha = \int_{\mathcal{E}(r_{\frac{\alpha}{C}})\cap\mathcal{S}} C\Pi(\varrho)\,\mathrm{d}\varrho. \tag{2.107}$$

Note that the right equation is exactly the defining Eq. (2.103) for the positive radius r_α^+ if $r_\alpha^+ = r_{\frac{\alpha}{C}}$.

Now, assume that a part of the ellipsoid $O = \mathcal{E}(r_{\frac{\alpha}{C}}) \setminus \mathcal{S} \neq \emptyset$ lies outside the psd states. Then, as can be seen on the right side of Fig. 2.3, we need to enlarge r_α^+ to compensate for the lost probability weight of O. The latter cannot be vanishing, since the Gaussian density $\Pi(\varrho)$ is strictly positive. Therefore, $r_\alpha^+ > r_{\frac{\alpha}{C}}$ in this case. □

Of course, the difference between $r_{\frac{\alpha}{C}}$ and r_α^+ may in general become too small to be efficiently detectable. However, we will show that for the instances of the number partition problem encoded in Problem 2.13, this is not the case. A first step toward this is the following Lemma.

Lemma 2.23. *Let $\boldsymbol{a} \in \mathbb{N}^d$ be an instance of the number partition problem 2.10 and*

$$\mathcal{E}_{\boldsymbol{a}} = \left\{ \varrho_0 + R_1 \sum_{i=1}^{i_d} u_i \sigma_i^+ R_2 \sum_{i=i_d+1}^{d^2-1} u_i \sigma_i \colon \boldsymbol{u}^T \boldsymbol{u} \leq 1 \right\} \tag{2.108}$$

the corresponding encoding ellipsoid for Problem 2.13 defined in Section 2.4.4. Then, there exists a polynomial $\tilde{p}$ such that if $\mathcal{E}_{\boldsymbol{a}}$ is not a subset of $\mathcal{S}$, there is an element $\varrho \in \mathcal{E}_{\boldsymbol{a}}$ with

$$\operatorname{mineig}(\varrho) \leq -\tilde{p}(\|\boldsymbol{a}\|)^{-1} < 0. \tag{2.109}$$

Proof. The main proof idea is to trace back the proof for polynomial gap in Lemma 2.17. Recall that Eqs. (2.64) and (2.67) ensure that if $\boldsymbol{a}$ has a partition, there is a $\boldsymbol{\Psi} \in \{\pm 1\}^d$ such that $\boldsymbol{a} \cdot \boldsymbol{\Psi} = 0$ and

$$d^2 - \sum_k \psi_k^4 + \left(d - \frac{(\boldsymbol{a} \cdot \boldsymbol{\psi})^2}{1 + \|\boldsymbol{a}\|^2} \right)^2 - C_2 (\boldsymbol{a} \cdot \boldsymbol{\psi})^4 = C_1 + p(\|\boldsymbol{a}\|)^{-1}. \tag{2.110}$$

By tracing back the steps which lead to this equation, we find for $|\Psi\rangle := \sum_{k=1}^d \psi_k / \sqrt{d} |k\rangle$

$$\frac{2(R_1^2 - R_2^2)}{d} p(\|\boldsymbol{a}\|)^{-1} + \langle \Psi | \varrho_0 | \Psi \rangle^2 \tag{2.111}$$

$$= R_1^2 \sum_i \left(\langle \Psi | \sigma_i^{(x)} | \Psi \rangle \right)^2 + R_2^2 \sum_i \left(\langle \Psi | \sigma_i^{(y,z)} | \Psi \rangle \right)^2 \tag{2.112}$$

$$=: \sum_i R_i^2 (\langle \Psi | \sigma_i | \Psi \rangle)^2 \tag{2.113}$$

Due to the special choice for ϱ_0 in (2.54) and $\boldsymbol{a} \cdot \boldsymbol{\psi} = 0$, we have

$$\langle \Psi | \varrho_0 | \Psi \rangle = \frac{q}{d} \tag{2.114}$$

with q defined in (2.57). Therefore, we can rewrite Eq. (2.111) as

$$\begin{aligned} \langle \Psi | \varrho_0 | \Psi \rangle - \sqrt{\sum_i R_i^2 \langle \Psi | \sigma_i | \Psi \rangle^2} \quad &= \frac{q}{d} \left(1 - \sqrt{1 + \frac{2d(R_1^2 - R_2^2)}{q^2\, p(\|\boldsymbol{a}\|)}} \right) \\ &\leq -\min \left(\frac{R_1^2 - R_2^2}{2q\, p(\|\boldsymbol{a}\|)}, \frac{2q}{d} \right) \end{aligned} \tag{2.115}$$

where we have used

$$1 - \sqrt{1 + x^2} \leq \begin{cases} -x^2/4 & x \leq 2\sqrt{2} \\ -2 & x > 2\sqrt{2} \end{cases} \tag{2.116}$$

Since all the constants on the right hand side of Eq. (2.115) can be expressed as polynomials in the input, it defines the polynomial $\tilde{p}(\|\boldsymbol{a}\|)$ of the lemma. The left hand side of Eq. (2.115) is equal to $\langle\Psi|\varrho|\Psi\rangle$, where

$$\varrho = \varrho_0 + \sum_i R_i u_i \sigma_i \in \mathcal{E}_{\boldsymbol{a}} \tag{2.117}$$

for the special choice of u from (2.44). The claim of the lemma follows for this ϱ using Eq. (2.115). □

We will now show how the explicitly parameterized ellipsoid (2.108) can be encoded as a MVCR-ellipsoid of a Gaussian distribution.

Lemma 2.24. *Denote by*

$$\mathcal{E}^* = \left\{ \varrho_0 + \sum_{i=1}^{d^2-1} u_i R_i \sigma_i \colon \|\boldsymbol{u}\|_2 = 1 \right\} \tag{2.118}$$

an ellipsoid $\mathcal{E}^* \subseteq \mathbb{H}$, which is axis-aligned with the coordinate axes defined by the generalized Pauli operators. Then, $\mathcal{E}^*$ can be encoded as a $\frac{\alpha}{C}$ MVCR-ellipsoid for a Gaussian distribution with mean $\varrho_0 \in \mathcal{S}$ and covariance matrix Σ. The latter is diagonal in the generalized Bloch basis σ_i with entries $\Sigma_{ij} = R_i^2 \delta_{ij}$ and for the corresponding radius we have $r_{\frac{\alpha}{C}} = \sqrt{2}$. Hence, the credibility is given by

$$\alpha = C\, P\left(\tfrac{N}{2}, 1\right), \tag{2.119}$$

which can be calculated efficiently with exponential precision for given C and N.

Proof. Since the generalized Pauli operators form an orthogonal system with $\mathrm{tr}(\sigma_i \sigma_j) = 2\delta_{ij}$, we find for $\varrho \in \mathcal{E}^*$

$$\|\varrho\|_2^2 = \sum_{i,j} u_i u_j\, R_i R_j \left(\Sigma^{-1}\right)_{ij} 2\delta_{ij} = 2\|\boldsymbol{u}\|_2^2. \tag{2.120}$$

Therefore, $\mathcal{E}^* = \mathcal{E}(\sqrt{2})$ with mean ϱ_0 and the stated covariance matrix. The efficient computation of the credibility (2.119) is given later in the proof of Lemma 2.26. □

Based on the gap proven in Lemma 2.23, we will now turn to the following question: In case Eq. (2.106) does not hold – that is the corresponding ellipsoid is not fully contained in the psd states – is the corresponding gap always large enough to be efficiently detectable?

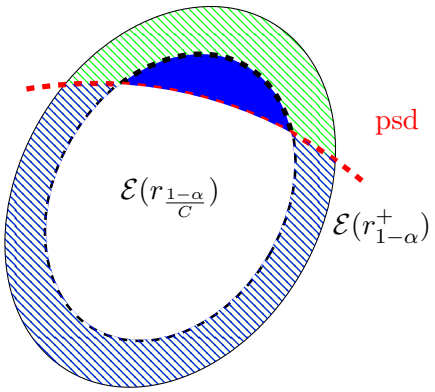


Figure 2.4.: Same as Fig. 2.3 (right). Note that the solid blue and hatched blue regions need to have the same volume.

Lemma 2.25. *Let $\boldsymbol{a} \in \mathbb{N}^d$ be an instance of the number partition problem and denote by $\mathcal{E}_{\boldsymbol{a}}$ the corresponding encoding ellipsoid as given by Eq.* (2.108)*. Furthermore, denote by $\Pi_{\varrho_0,\Sigma}$ the Gaussian density, which encodes $\mathcal{E}_{\boldsymbol{a}} = \mathcal{E}(r_{\frac{\alpha}{C}})$ as an $\frac{\alpha}{C}$ credible region as given by Lemma 2.24. Assume that $\boldsymbol{a}$ has a balanced sum partition and, therefore, $\mathcal{E}_{\boldsymbol{a}}$ is not a subset of $\mathcal{S}$.*

Then, there exists a polynomial p such that

$$r_{\alpha}^{+2} - r_{\frac{\alpha}{C}}{}^{2} \geq 2^{-p(\log \|\boldsymbol{a}\|_1)}. \tag{2.121}$$

Here, $\|\boldsymbol{a}\|_1 = \sum_k |a_k|$. In words, the gap of violation of Eq. (2.106) *can only become polynomially small in the logarithm of the size of the problem specification.*

Proof. First, let us lower bound the volume of $\mathcal{E}(r_{\frac{\alpha}{C}})$ that lies outside the psd states (the solid blue region in Fig. 2.4). From Lemma 2.23 we know, that there exists a $\varrho \in \mathcal{E}(r_{\frac{\alpha}{C}})$ with smallest eigenvalue smaller than $-\tilde{p}(\|\boldsymbol{a}\|)^{-1}$ for some polynomial $\tilde{p}$. This also gives us a lower bound on

$$\operatorname{dist}(\varrho, \mathcal{S}) = \inf_{\varrho' \in \mathcal{S}} \|\varrho - \varrho'\|_2. \tag{2.122}$$

From [Bha13, Theorem III.2.8] we know that for every $\varrho_+ \in \mathcal{S}$ the following bound holds:

$$\begin{aligned} \|\varrho - \varrho_+\|_2 &\geq \|\varrho - \varrho_-\|_\infty \geq \|\boldsymbol{\lambda}^{\uparrow}(\varrho) - \boldsymbol{\lambda}^{\uparrow}(\varrho_+)\|_2 \\ &\geq |\operatorname{mineig}(\varrho) - \operatorname{mineig}(\varrho_+)| \geq \tilde{p}(\|\boldsymbol{a}\|)^{-1}. \end{aligned} \tag{2.123}$$

Here, $\boldsymbol{\lambda}^{\uparrow}(\rho)$ denotes the vector of eigenvalues of ρ in ascending order. Therefore,

$$\operatorname{dist}(\varrho, \mathcal{S}) \geq \tilde{p}(\|\boldsymbol{a}\|)^{-1}. \tag{2.124}$$

This allows us to lower bound the volume of $\mathcal{E}(r_{\frac{\alpha}{C}})$ that lies outside the psd states by an ellipsoid with the same covariance, but radius $\left(2\,\tilde{p}(\|\boldsymbol{a}\|)\,\operatorname{maxeig}(\Sigma)\right)^{-1}$

$$\operatorname{Vol}\left(\mathcal{E}(r_{\frac{\alpha}{C}}) \setminus \mathcal{S}\right) \geq \frac{\pi^{\frac{N}{2}}\,|\Sigma|}{\Gamma(\frac{N}{2}+1)} \frac{1}{\left(2\tilde{p}(\|\boldsymbol{a}\|)\,\operatorname{maxeig}(\Sigma)\right)^{N}} \tag{2.125}$$

$$\tag{2.126}$$

Furthermore, we have

$$\operatorname{Vol}\left(\mathcal{E}(r^{+}_{1-\alpha}) \setminus \mathcal{E}(r_{\frac{1-\alpha}{C}})\right) = \operatorname{Vol}\left(\mathcal{E}(r_{\frac{1-\alpha}{C}}) \setminus \mathcal{S}\right) \tag{2.127}$$

since the solid blue and hatched blue regions in Fig. 2.4 must be of same size. We now relate the volume inequality (2.125) to a lower bound for the mass of the ellipsoid outside the psd states w.r.t. the Gaussian density: Due to the set of states $\mathcal{S}$ having finite radius $\sqrt{\frac{2(d-1)}{d}}$ [Kim03, Eq. (18)], we must have $r^{+}_{\alpha} \leq 2\sqrt{2}$. Therefore,

$$P\left(\tfrac{N}{2}, \tfrac{r^{+2}_{\alpha}}{2}\right) - P\left(\tfrac{N}{2}, \tfrac{r_{\frac{\alpha}{C}}{}^{2}}{2}\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\,|\Sigma|^{\frac{1}{2}}} \int_{\mathcal{E}(r^{+}_{\alpha}) \setminus \mathcal{E}(r_{\frac{\alpha}{C}})} \mathrm{e}^{-\frac{1}{2}\|\varrho-\varrho_0\|^2} \mathrm{d}^N\varrho \tag{2.128}$$

$$\geq \frac{\mathrm{e}^{-4}}{(2\pi)^{\frac{N}{2}}\,|\Sigma|^{\frac{1}{2}}} \operatorname{Vol}\left(\mathcal{E}(r^{+}_{\alpha}) \setminus \mathcal{E}(r_{\frac{\alpha}{C}})\right) \tag{2.129}$$

$$\geq \frac{\mathrm{e}^{-4}\pi^{\frac{N}{2}}\,|\Sigma|^{\frac{1}{2}}}{2^{\frac{N}{2}}\Gamma(\frac{N}{2}+1)} \frac{1}{\left(2\tilde{p}(\|\boldsymbol{a}\|)\,\operatorname{maxeig}(\Sigma)\right)^{N}} \tag{2.130}$$

$$=: 2^{-p(\log\|\boldsymbol{a}\|_1)-1} \tag{2.131}$$

Finally, note that the following crude inequality

$$P\left(\tfrac{N}{2}, \tfrac{r^{+2}_{\alpha}}{2}\right) - P\left(\tfrac{N}{2}, \tfrac{r_{\frac{\alpha}{C}}{}^{2}}{2}\right) = \int_{y}^{x} \frac{t^{\frac{N}{2}-1}\mathrm{e}^{-t}}{\Gamma(\frac{N}{2}+1)}\,\mathrm{d}t \leq x - y \tag{2.132}$$

holds for $x \geq y$, since the integrand is less than 1. Therefore, with Eq. (2.131)

$$r^{+2}_{\alpha} - r_{\frac{\alpha}{C}}{}^{2} \geq 2^{-p(\log\|\boldsymbol{a}\|_1)}, \tag{2.133}$$

which proofs the claim. □

We now turn to the problem of computing the normalization constant C for the restricted Gaussian distribution (2.100). First, we efficiently compute a credibility $\alpha' \in [0,1]$ such that the corresponding credible ellipsoid $\mathcal{E}(r_{\frac{\alpha'}{C}})$ is guaranteed to be contained in the psd states without knowing the value of C. This allows us to leverage Eq. (2.106) to compute C.

Lemma 2.26. *Let $\boldsymbol{a} \in \mathbb{N}^d$ be an instance of the number partition problem and denote by $\mathcal{E}_{\boldsymbol{a}}$ the corresponding encoding ellipsoid as defined by Eq.* (2.108)*. Denote by $\Pi_{\varrho_0,\Sigma}$ the Gaussian density, which encodes $\mathcal{E}_{\boldsymbol{a}}$ as an α credible region according to Lemma 2.24. Then, the ellipsoid $\mathcal{E}(r)$ is fully contained in the psd states provided*

$$r \leq \sqrt{\frac{d}{2(d-1)}} \frac{\operatorname{mineig} \varrho_0}{\sqrt{\operatorname{maxeig} \Sigma}} \tag{2.134}$$

Proof. We know that for any $\varrho \in \mathcal{E}(r)$ with r fulfilling (2.134) the following inequalities hold

$$\begin{aligned} \|\varrho - \varrho_0\| &\leq \frac{1}{\sqrt{\operatorname{mineig} \Sigma^{-1}}} \|\varrho - \varrho_0\|_\Sigma \\ &\leq \frac{1}{\sqrt{\operatorname{mineig} \Sigma^{-1}}} r \\ &\leq \sqrt{\frac{d}{2(d-1)}} \operatorname{mineig} \varrho_0. \end{aligned}$$

Here, we have used $\operatorname{mineig} \Sigma^{-1} = (\operatorname{maxeig} \Sigma)^{-1}$, which holds for any positive definite matrix Σ. Therefore, $\mathcal{E}(r) \subseteq \mathcal{S}$ due to Lemma 2.18. □

Lemma 2.27. *Using the same notation as Lem. 2.26 and assuming Prob. 2.20 can be solved efficiently. Then, for every instance $\boldsymbol{a}$ of the number partition problem consider the corresponding ellipsoid encoding distribution according to Lemma 2.24 with parameters θ, Σ. Then, we can efficiently approximate the normalization constant C of $\Pi^+_{\theta,\Sigma}$ with exponentially small multiplicative error. More precisely, we have*

$$C = \tilde{C}(1+\epsilon), \tag{2.135}$$

where $\tilde{C}$ can be computed in polynomial time making the correction term ϵ exponentially small.

Proof. Due to Lemma 2.26 and $\operatorname{mineig} \theta > 0$, we can always find an $r > 0$ such that $\mathcal{E}(r)$ is fully contained in the psd. Indeed, the eigenvalues of θ and Σ are readily calculated because of their particular simple form in Eq. (2.54) and Lemma 2.24:

$$\sqrt{\frac{d}{2(d-1)}} \frac{\operatorname{mineig} \theta}{\sqrt{\operatorname{maxeig} \Sigma}} = \frac{q}{R_1 \sqrt{2d(d-1)}} \tag{2.136}$$

Set[3]

$$\alpha := P\left(\tfrac{N}{2}, \tfrac{r^2}{2}\right). \tag{2.137}$$

Since we can choose r as small as we want, we may assume that $x = \frac{r^2}{2} \ll 1 < \frac{N}{2}$. In this regime, we can expand the normalized incomplete Γ-function P in a power series [GST12]

$$P\left(\tfrac{N}{2}, x\right) = \frac{x^{\frac{N}{2}} \mathrm{e}^{-x}}{\Gamma\left(\frac{N}{2}+1\right)} \sum_{k=0}^{\infty} \frac{x^k}{\left(\frac{N}{2}+1\right)_k}, \tag{2.138}$$

where

$$\left(\tfrac{N}{2}+1\right)_k = \frac{\Gamma\left(\frac{N}{2}+k+1\right)}{\Gamma\left(\frac{N}{2}+1\right)}. \tag{2.139}$$

Truncating the series in Eq. (2.138) for $k \geq k_0$

$$P\left(\tfrac{N}{2}, x\right) = P_{k_0}\left(\tfrac{N}{2}, x\right) + R_{k_0}\left(\tfrac{N}{2}, x\right), \tag{2.140}$$

with

$$P_{k_0}\left(\tfrac{N}{2}, x\right) = \frac{x^{\frac{N}{2}} \mathrm{e}^{-x}}{\Gamma\left(\frac{N}{2}+1\right)} \sum_{k=0}^{k_0} \frac{x^k}{\left(\frac{N}{2}+1\right)_k} \tag{2.141}$$

we can derive a bound on the truncation error $R_{k_0}(\frac{N}{2}, x)$ [GST12, Eq. (2.18)]

$$R_{k_0}(\tfrac{N}{2}, x) \leq \frac{x^{\frac{N}{2}+k_0} \mathrm{e}^{-x}}{\Gamma(\frac{N}{2}+k_0+1)} \frac{\frac{N}{2}+k_0}{\frac{N}{2}+k_0-x-1}. \tag{2.142}$$

Since $x \ll 1$, the term x^{k_0} ensures that we can make the error in computing α exponentially small using only polynomial time in evaluating $P_{k_0}(\frac{N}{2}, x)$.

Now, assume that we have computed $\tilde{\alpha} = \alpha - \epsilon$ for some truncation error $\epsilon = R_{k_0}(\frac{N}{2}, x) > 0$. We may now use the postulated efficient algorithm for Prob. 2.20 to compute the radius of the manifestly positive MVCR $r^+_{\tilde{\alpha}}$ and, hence, using Eq. (2.106) the normalization constant: Since $C > 1$, we have with $r_\alpha = r$

$$r_{\frac{\tilde{\alpha}}{C}} = r_{\frac{\alpha-\epsilon}{C}} < r_\alpha \implies \mathcal{E}(r_{\frac{\tilde{\alpha}}{C}}) \subseteq \mathcal{S} \implies r_{\frac{\tilde{\alpha}}{C}} = r^+_{\tilde{\alpha}} \leq r_\alpha. \tag{2.143}$$

Therefore, the ellipsoid with radius $r^+_{\tilde{\alpha}}$ is also contained in the psd states. The same holds true if we replace $r^+_{\tilde{\alpha}}$ by the actual output $r^+_{\tilde{\alpha}} \pm \delta$ of the postulated efficient algorithm for Prob. 2.19 Here, δ denotes the accuracy that is part of the input to the

[3] Note that α does not denote the credibility used for encoding the ellipsoid in question, but an auxiliary ellipsoid used for computing C here.

problem of computing $r_{\tilde{\alpha}}^+$. By choosing δ small enough and possibly replacing the original radius r by $r - \delta$, we can ensure that

$$\mathcal{E}(r_{\tilde{\alpha}}^+ \pm \delta) \subseteq \mathcal{S}, \tag{2.144}$$

as well. Therefore, Eq. (2.106) holds and we find

$$\frac{\tilde{\alpha}}{C} = P\left(\tfrac{N}{2}, \tfrac{{r_{\tilde{\alpha}}^+}^2}{2}\right) \tag{2.145}$$

$$= P\left(\tfrac{N}{2}, \tfrac{(r_{\tilde{\alpha}}^+ \pm \delta)^2}{2}\right) - \frac{1}{\Gamma(\frac{N}{2})} \int_{\frac{{r_{\tilde{\alpha}}^+}^2}{2}}^{\frac{(r_{\tilde{\alpha}}^+ \pm \delta)^2}{2}} t^{\frac{N}{2}-1} \mathrm{e}^{-t} \mathrm{d}t. \tag{2.146}$$

The first addend on the right hand side can be evaluated using the same series expansion as in Eq. (2.140), since we are in the same regime $\frac{{r_{\tilde{\alpha}}^+}^2}{2} \ll \frac{N}{2}$. The second addend can be bounded by

$$\left| \frac{1}{\Gamma(\frac{N}{2})} \int_{\frac{{r_{\tilde{\alpha}}^+}^2}{2}}^{\frac{(r_{\tilde{\alpha}}^+ \pm \delta)^2}{2}} t^{\frac{N}{2}-1} \mathrm{e}^{-t} \mathrm{d}t \right| < \frac{\left(2 r_{\tilde{\alpha}}^+ \delta + \delta^2\right)}{2} \tag{2.147}$$

since

$$\frac{t^{\frac{N}{2}-1} \mathrm{e}^{-t}}{\Gamma(\frac{N}{2})} < 1. \tag{2.148}$$

Let us assume w.l.o.g. $r_{\tilde{\alpha}}^+ \leq 1$. This bound, as well as the error bound $\epsilon' > 0$ for the finite series-evaluation of P in (2.145) leads to

$$\frac{\tilde{\alpha}}{C} = P_{k_0}\left(\tfrac{N}{2}, \tfrac{(r_{\tilde{\alpha}}^+ \pm \delta)^2}{2}\right) + \epsilon' \pm D\delta \tag{2.149}$$

for some appropriate constant D. A little arithmetic gives

$$C = \frac{\tilde{\alpha}}{P_{k_0}(\ldots)} \left(1 - \frac{\epsilon' \pm D\delta}{P_{k_0}(\ldots) + \epsilon' \pm D\delta}\right). \tag{2.150}$$

By assumption we can make both ϵ' and δ exponentially small using only polynomial time Furthermore, $P_{k_0}(\frac{N}{2}, x) \uparrow P(\frac{N}{2}, x)$ for $k_0 \to \infty$ and the correction to

$$\tilde{C} = \frac{\tilde{\alpha}}{P_{k_0}\left(\tfrac{N}{2}, \tfrac{(r_{\tilde{\alpha}}^+ \pm \delta)^2}{2}\right)} \tag{2.151}$$

in Eq. (2.150) can be made exponentially small using polynomial time. On the other hand, $\tilde{C}$ can be computed in polynomial time as well. □

We now have all the necessary parts for the proof of the main theorem 2.21, which concludes this section.

Proof of Thm. 2.21. The proof follows the outline stated in the main text: First, we encode the ellipsoid of Problem 2.13 to be checked as a MVCR of a Gaussian with mean ϱ_0 and covariance matrix Σ according to Lemma 2.24. Using Lemma 2.27, we compute an estimate $\tilde{C}$ to the normalization constant C. Using the techniques from the proof of the aforementioned Lemma, we may compute an estimate

$$\alpha = C\, P\left(\tfrac{N}{2}, 1\right) = \tilde{C}(1+\epsilon)\left(P_{k_0}\left(\tfrac{N}{2}, 1\right) + \epsilon'\right) = \tilde{\alpha} + \epsilon''. \tag{2.152}$$

This can be done for exponential small errors ϵ, ϵ' in polynomial time. Here, the computable value is given by

$$\tilde{\alpha} = \tilde{C}\, P_{k_0}\left(\tfrac{N}{2}, 1\right). \tag{2.153}$$

An exponential small difference of α and $\tilde{\alpha}$ also implies an exponential small difference of $r^+_{1-\alpha}$ and $r^+_{\tilde{\alpha}}$: Set $x := r^+_\alpha$ and $\tilde{x} := r^+_{\tilde{\alpha}}$ and assume $x > \tilde{x}$ – the opposite case can be treated along the same lines by choosing a larger constant as a bound for $\tilde{x}$. Following Eq. (2.131), we have

$$\begin{aligned} P\left(\tfrac{N}{2}, \tfrac{x^2}{2}\right) - P\left(\tfrac{N}{2}, \tfrac{\tilde{x}^2}{2}\right) &\geq \frac{\mathrm{e}^{-4}}{(2\pi)^{\frac{N}{2}}\,|\Sigma|^{\frac{1}{2}}} \operatorname{Vol}\left(\mathcal{E}(x) \setminus \mathcal{E}(\tilde{x})\right) \\ &= \frac{\mathrm{e}^{-4}}{2^{\frac{N}{2}}\,\Gamma(\frac{N}{2}+1)}\left(x^N - \tilde{x}^N\right). \end{aligned}$$

Since for fixed N, the left hand side can be made exponentially small in polynomial time by improving $\tilde{\alpha}$, so can the right hand side. Therefore, the difference $|x - \tilde{x}|$ can be made exponentially small as well.

Now, choose the errors ϵ and ϵ' in such a way that

$$|r^+_\alpha - r^+_{\tilde{\alpha}}| \leq \frac{\Delta}{4}. \tag{2.154}$$

Here, $\Delta = 2^{-p(\log \|\boldsymbol{a}\|_1)}$ is the (at worst exponentially small) gap from Lemma 2.25. Furthermore, we run the algorithm for computing $r^+_{\tilde{\alpha}}$ with precision $\delta = \frac{\Delta}{4}$ and denote the result by $\tilde{r}$. If $|\tilde{r} - \sqrt{2}| \leq \frac{\Delta}{2}$, we know that $r^+_\alpha = r_{\frac{\alpha}{C}}$ and the ellipsoid is fully contained in the psd states. Otherwise we know that it is not. □

2.6. Conclusion & outlook

The goal of this work is to provide an absolute "upper bound" on what we can expect from algorithms computing error regions for QSE and to demonstrate that there is

a trade-off between optimality and efficiency. This work should not be understood as providing a no-go theorem for efficient algorithms in practice. As discussed in the end of Section 2.2, the negative result of this work does not rule out efficient algorithms for practically acceptable approximations to optimal regions. Also, there is no indication that the various approaches used in practice give rise to regions that are far from optimal or do not have the advertised coverage. The reason our result leaves room for feasible approaches in practice are twofold: First, like any result showing **NP**-hardness, we prove that there is no efficient algorithm solving the exact problem deterministically for any instance. Hence, our result neither precludes the existence of efficient approximate or probabilistic algorithms, nor cannot make any statement about average case hardness. Second, although the experimental effort necessary for full-fledged QSE scales polynomially in the dimension of the system – and is, therefore, efficient in the sense of computational complexity – in practice other characterization techniques such as randomized benchmarking or direct fidelity estimation become more important for larger dimensions. It should now be the goal of future work to further close down the gap between existing positive results and the proven no-go theorems from either side.

More specifically, due to the simplifying assumptions made, we investigate computational intractability that is solely caused by the quantum constraints and not by the general complications in high-dimensional statistics. In the Bayesian settings we show that minimal volume (w.r.t. the Hilbert-Schmidt measure) credible regions for truncated Gaussian posterior distributions are hard to compute. Therefore, the problem of determining MVCR for QSE cannot be solved efficiently as well, since any algorithm solving the latter must also be able to solve instances with the specific prior used in Prob. 2.20.

The result for frequentist confidence regions is somewhat weaker since optimal confidence regions for high-dimensional Gaussian distributions are not known for most natural notions of optimality. Nevertheless, Gaussian confidence ellipsoids constitute a viable choice due to their simplicity and tractability. However, our results show that the constraints imposed by quantum mechanics render the task of characterizing the confidence regions for the constrained problem computationally intractable – even under the simplifying assumptions made. Of course, any more general setting encompassing the Gaussian approximation will be at least as hard to treat as the one used in this work. Furthermore, it also shows that computing any confidence region estimator yielding ellipsoids when the constraints are not active (and anything possibly better when they are) involves solving **NP**-hard problems.

Recently, the mathematical statistics community has started to analyze the trade-offs between computational complexity and optimality in inference problems – see e.g. [BR13a; BR13b; ZWJ14]. Early papers concentrated on the problem of *sparse principal component analysis*, which roughly asks whether the covariance matrix of

a random vector possess a sparse eigenvector with large eigenvalue [BR13a; BR13b; ZWJ14]. Later works have addressed the much better-studied problem of sparse inference [ZWJ14]. The main difference between these papers and the present one is that we always condition on a data set and show that certain operations for quantifying uncertainty given the data are hard. This approach is canonical for a Bayesian analysis, but merely "natural" for frequentist confidence regions (c.f. Section 2.1.1). In contrast, Refs. [BR13a; BR13b; ZWJ14] analyze the "global" performance of orthodox estimators – i.e. they do not require looking at worst-case scenarios over the data. References [BR13a; BR13b; ZWJ14] achieve this by reducing a certain problem ("hidden clique") – that is conjectured to be hard in the average case – to the sparse PCA problem; while [ZWJ14] employs a more subtle argument involving the non-uniform complexity class $\mathbf{P}/\mathrm{poly}$. It would be very interesting to adapt such arguments to the problem of quantum uncertainty quantification.

Of course, from the practical point of view, "positive" results – i.e. new algorithms to solve the problem – would be more beneficial. Here, recent work on sampling distributions restricted to convex bodies [CV14; CV15] could be a starting point for further investigations.

Beside quantum state tomography, our results might also be relevant to problems involving psd constraints such as the estimation of covariance matrices.

3. Characterizing linear-optical networks via PhaseLift

Even though photonics is often considered the "ugly duckling" of the approaches to quantum computing and simulation [Rud17], it has two main advantages over other approaches: It is inherently robust towards stochastic noise and integrated photonic devices can be fabricated using present-day fabrication techniques for silicon-based semi-conductors [Rud17]. Passive and reconfigurable linear optical circuits have been proposed and demonstrated for many applications including telecommunications [Mil15], machine learning [She+17] as well as quantum computation [Car+15a] and simulation [Har+17]. With the continuing development of large-scale integrated photonic platforms [Sil+16; Seo+16], practical and reliable techniques for characterizing and validating the operation of these devices are crucial. Here, characterization refers to the problem of recovering a full mathematical description of the linear optical devices in terms of its *transfer matrix* M from measurable quantities.

In this chapter, we propose an efficient, robust, and conceptually simple technique for characterizing linear optical circuit by exploiting a connection to the phase retrieval problem [Wal63]. Not only do we adapt existing results from phase retrieval and low-rank matrix recovery to the problem of characterizing linear-optical networks, we also propose a measurement ensemble tailored to this specific application. To present the rigorous analysis of both approaches, we develop a unified proof strategy. Besides having these stringent recovery guarantees, the *PhaseLift* reconstruction algorithm proposed here is robust to noise and efficient with respect to the number of measurements.

Well-known techniques for characterizing linear optical circuits include quantum process tomography with non-classical [OBr+04] or coherent [Rah+11] states, though the sample complexity of these approaches scales exponentially with the number of modes. Simpler protocols tailored to linear optics have been proposed that use either single and two-photon probe states [LO12; Dha+16; Spa+17] or multimode coherent states [Rah+13; TSW16]. The most similar scheme to the one presented here is [Rah+13], where coherent light is input into single modes and split over pairs of modes with the intensity at each output measured. While their recovery method is strikingly simple and relies only on $2n - 1$ input configurations, for each configuration it requires varying over a phase shift between the two modes until maximal constructive interference is observed. Hence, the experiment needs to be performed

interactively, where the phase shift is adjusted gradually throughout an individual measurement, or a large number of phase shifter settings need to be probed. Both alternatives require a large number of measurements to be performed in order to recover M successfully. Furthermore, by construction, the protocol from [Rah+13] utilizes the obtained reconstruction of the first row to recover the remaining rows of M. This makes it a priori susceptible towards noise as any error in the determination of the first row propagates to the remaining rows.

This chapter is structured as follows: In Section 3.1, we recapitulate the fundamental problem of characterizing linear optical devices using either coherent states of light or single photon states. Section 3.2 is concerned with introducing the fundamental problem of phase retrieval. Section 3.3 contains the main theoretical results of this work, namely the proposed measurement scheme, the related recovery guarantees for the PhaseLift reconstruction algorithm as well as the characterization of linear-optical networks via PhaseLift. We present results from numerical and experimental investigation in Section 3.4. Section 3.5 concludes this chapter and provides an outlook on possible future work.

3.1. Device characterization

Mathematically, a linear optical device is fully characterized by its *transfer matrix*, which relates the output to the input of the device by

$$a^\dagger{}_j \to b^\dagger{}_j = \sum_i M_{i,j} a^\dagger{}_i. \tag{3.1}$$

Here, $a^\dagger{}_j$ and $b^\dagger{}_j$ denote the creation operators of the j-th input and output mode, respectively. Determining M experimentally is the crucial step to validate and verify a linear optical circuit. For this purpose, we propose a protocol that can be implemented easily in an experiment using either classical laser light or single photon sources. We first introduce the former approach as it is conceptually simpler.

For now, we assume that the input is described by a classical multi-mode coherent state $|\alpha\rangle = |\alpha_1, \ldots, \alpha_n\rangle$ with

$$|\alpha\rangle = \mathrm{e}^{-\frac{\|\alpha\|_{\ell_2}}{2}} \sum_{k_1,\ldots,k_n} \frac{\alpha_1^{k_1} \ldots \alpha_n^{k_n}}{\sqrt{k_1! \ldots k_n!}} |k_1\rangle \ldots |k_n\rangle. \tag{3.2}$$

Then, due to Eq. (3.1), the output is a coherent state $|\beta\rangle$ as well and its components are given by

$$\beta_j = \sum_k M_{j,k}\, \alpha_k. \tag{3.3}$$

Note that for an ideal, unitary transfer matrix, $\|\alpha\|_{\ell_2} = \|\beta\|_{\ell_2}$ as the squared norm of a coherent state vector describes its total intensity. The standard measurable quantities in an optical experiment with coherent states are the *intensities* of the output modes

$$I_j(\alpha) = |\beta_j|^2 + \epsilon_j = \left| \sum_k M_{j,k}\, \alpha_k \right|^2 + \epsilon_j \tag{3.4}$$

for certain coherent inputs $|\alpha\rangle$. Here, ϵ_j describes noise due to statistical fluctuations or systematic errors. A schematic of such an experiment is depicted in Fig. 3.1 a). Although the output coherent states (3.3) are linear in M, the resulting intensity measurements (3.4) are quadratic in M and oblivious to the phases of β. Therefore, the problem of reconstructing M from such measurements is ill-posed and requires deliberate utilization of interference between the modes to recover the phases of M.

We propose an approach for recovering M from the measurements (3.4) with the coherent states α sampled randomly from appropriate distributions. Preparing theses states reliably is the major challenge of implementing the proposed protocol experimentally. A first experimental demonstration is performed using the universal linear optics device from [Car+15a]: The silica-on-silicon device performs a linear-optical circuit comprising 30 directional couplers and 30 tunable thermo-optic phase-shifters on six optical waveguides. Using the setup outlined in Fig. 3.1, we are able to prepare any coherent input state from a single laser input in the bottom mode using the left-most cascade of couplers and phase-shifters. The remaining triangular array of components colored blue in Fig. 3.1 is then sufficient to implement any five mode unitary transfer matrix M [Rec+94]. Reconfiguring the target M then enables us to experimentally test the protocol across a number of configurations including Identity, Swap, and Fourier matrices as well as Haar random unitaries.

To prepare an arbitrary coherent state $|\tilde{\alpha}\rangle$ experimentally, we use the red colored cascade of couplers and phase-shifters in Fig. 3.1: First, we input a single laser with intensity $\|\tilde{\alpha}\|^2$ in the first mode, which is mathematically described by the coherent state $|\|\tilde{\alpha}\|e_1\rangle$. Here, e_1 denotes the first canonical basis vector. Then, the couplers and phase-shifters in the left-most cascade are set in such a way to implement a transfer matrix $P(\alpha)$ with

$$\tilde{\alpha} = P(\alpha)(\|\tilde{\alpha}\|e_1). \tag{3.5}$$

Note that due to linearity, P can only depend on the normalized coherent label $\alpha = \frac{\tilde{\alpha}}{\|\tilde{\alpha}\|}$. Therefore, for the proposed preparation scheme, it is beneficial to consider ensembles of coherent states with fixed norm as it allows for keeping the input laser at constant intensity. Alternatively, we could redirect parts of the light in the topmost, unobserved mode, which is challenging when the required $\|\tilde{\alpha}\|$ varies a lot.

Although performing recovery of M using only classical sources of light and photodiodes simplifies the experiment, it also has a large drawback in practice: Our main motivation for studying linear optical devices is their application in quantum computing, which requires the use of single-photon sources. However, readily available laser and single photon sources often have slightly different characteristics such as wavelength or polarization. Since the properties of the components, and therefore, also the transfer matrix are generally dependent on these characteristics, a characterization using coherent light is generally unsuitable for predicting the performance of the device when used with single photon sources. Instead, the device should be evaluated under the same experimental conditions under which it will be used. For this purpose, we now turn to an experimental implementation of the idea introduced above based on single-photon sources and detectors.

The idea is to estimate the outcomes of the intensity measurements (3.4) using single photon states: If we set the preparation stage to $P(\alpha)$, but feed a single photon Fock state in the bottom waveguide, the prepared state prior to M is

$$|\psi(\alpha)\rangle = \sum_j \alpha_j {a_j}^\dagger |\mathrm{o}\rangle, \tag{3.6}$$

where $|\mathrm{o}\rangle$ denotes the vacuum state. For $|\psi(\alpha)\rangle$ to be well-normalized, we need to choose $\|\alpha\|_{\ell_2} = 1$. The probability of measuring the photon at detector j is then given by

$$p_j = \mathbb{P}(j|\alpha) = \left| \sum_k M_{j,k}\alpha_k \right|^2 . \tag{3.7}$$

Hence, finite-sample frequency estimates of the probabilities (3.7) are equivalent to the noisy intensity measurements (3.4). Note that α is now simply a parameter vector for the achievable single-photon Fock states (3.6).

To estimate the probabilities (3.7), we subsequentially feed N single photon Fock states into the device such that they do not interfer with each other. Then, the photon counting statistics is governed by a multinomial distribution, i.e.

$$\mathbb{P}(N_1, \ldots, N_n|\alpha) = \frac{N!}{N_1! \ldots N_n!} p_1^{N_1} \times \cdots \times p_n^{N_n}\, \delta_{N_1+\cdots+N_n,N}. \tag{3.8}$$

Here the right hand side is the probability of simultaneously measuring N_j photons in the j-th mode for $j = 1, \ldots, n$.

3.2. Phase retrieval

The crucial observation of this work is that measurements (3.4) closely resemble the model of the *phase retrieval problem*, i.e. the problem of recovering a complex vector

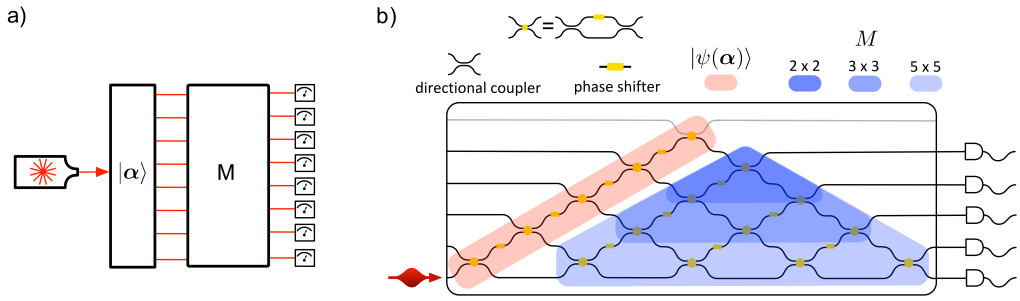


Figure 3.1.: Schematic of PhaseLift characterization protocol and experiment. a) Protocol summary using coherent states: A calibrated and trusted optical network is used to prepare multimode coherent states $|\alpha\rangle$, sampled from an appropriate ensemble. These states are then fed into the unknown linear optical device described by the transfer matrix M, and the intensities at each output mode are measured. b) Experimental implementation using single photon sources: Heralded single photons are injected into the bottom waveguide of a six-mode integrated photonic device. A cascade of Mach-Zehnder interferometers is used to prepare single-photon states $|\psi(\alpha)\rangle$ over the bottom five modes of the device. The remainder of the device is used to implement arbitrary 2, 3 and 5 dimensional unitary transformations which are to be characterized. Each output port is coupled to a single photon detector.

$x \in \mathbb{C}^n$ from m intensity measurements of the form

$$y^{(l)} = \left|\langle \alpha^{(l)}, x\rangle\right|^2 + \epsilon^{(l)} \quad l = 1, \ldots, m. \tag{3.9}$$

Here, $\alpha^{(l)} \in \mathbb{C}^n$ denote measurement vectors and $\epsilon^{(l)}$ the additive measurement errors. The major difficulty in recovering x from these intensity measurements is the loss of phase information. In order to infer the phase information, we need to exploit interference effects by carefully selecting different measurement vectors. Note, x can only be recovered up to a global phase since x and $\mathrm{e}^{\mathrm{i}\phi}x$ are indistinguishable from the measurements (3.9) for any phase angle ϕ.

One practical solution to the phase retrieval problem is based on its connection to the field of low-rank matrix recovery. The quadratic measurements of x in Eq. (3.9) can be rewritten as

$$\left|\langle x, \alpha^{(l)}\rangle\right|^2 = \mathrm{tr}\left((|\alpha^{(l)}\rangle\langle\alpha^{(l)}|)(|x\rangle\langle x|)\right). \tag{3.10}$$

This "lifts" the phase retrieval problem to the problem of recovering the positive semi-definite (psd) rank-1 matrix $|x\rangle\langle x|$ from linear measurements. Note that an efficient solution to this problem needs to exploit the low-rank constraint, as we have embedded the low-complexity signal $|x\rangle\langle x|$ into a n^2 dimensional ambient space. This problem – and its generalization to arbitrary low-rank matrices – has been studied extensively in the field of *low-rank matrix recovery*, see e.g. [ARR14; CR09; CP11; RFP10; Gro11; Che15] for a highly incomplete list of references.

The fundamental idea is to find the matrix Z with the smallest rank that is compatible with the observations. As an example, consider the idealized noiseless case of Eq. (3.9), i.e. $\epsilon^{(l)} = 0$. Then, we can reconstruct $|x\rangle\langle x|$ using the following rank-minimization problem

$$\begin{aligned} \underset{Z}{\text{minimize}} \quad & \mathrm{rank}\, Z \\ \text{subject to} \quad & \mathrm{tr}\left(|\alpha^{(l)}\rangle\langle\alpha^{(l)}|\, Z\right) = y_l \quad (l = 1, \ldots, m) \end{aligned} \tag{3.11}$$

provided the $\alpha^{(l)}$ suffice to single out $|x\rangle\langle x|$. However, rank minimization is **NP**-hard in general [BV04], and therefore, Eq. (3.11) cannot be solved efficiently. Nevertheless, there are algorithms for recovering $|x\rangle\langle x|$ that are computationally efficient with only a slight overhead in the sample complexity. Here, we consider the following convex algorithm termed *PhaseLift* [CSV13]

$$\begin{aligned} \underset{Z}{\text{minimize}} \quad & \sum_{l=1}^{m} \left|\mathrm{tr}\left(|\alpha^{(l)}\rangle\langle\alpha^{(l)}|\, Z\right) - y^{(l)}\right| \\ \text{subject to} \quad & Z \geq 0. \end{aligned} \tag{3.12}$$

From the minimizer $Z^\sharp$ of Eq. (3.12), we obtain the recovered signal vector $x^\sharp$ as follows: Consider the eigenvalue decomposition of $Z^\sharp$

$$Z^\sharp = \sum_i \lambda_i |z_i\rangle\langle z_i| \tag{3.13}$$

with $\|z_i\|_{\ell_2} = 1$ and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$. Then, we set

$$x^\sharp = \sqrt{\lambda_1} z_1. \tag{3.14}$$

Several analytic proofs of convergence have been established for phase retrieval via PhaseLift. With few notable exceptions [Kec16], these are probabilistic in nature and assume that each measurement vector is chosen from an appropriate distribution. Probabilistic in this context means that we allow for a small probability w.r.t. the random sampled measurement vectors of failing to reconstruct x. Paradigmatic examples are the Gaussian and the uniform (spherical) measurement ensemble [CSV13]: In case of the Gaussian ensemble, the components $\alpha_i^{(l)}$ of the measurement vectors are i.i.d. standard complex Gaussian random variables. In the uniform scheme, the $\alpha^{(l)}$ are chosen uniformly from the complex unit sphere. The two are closely related, as the latter arises by normalizing all Gaussian vectors to a fixed length. However, these two measurement ensembles are often too demanding for practical applications, which led to a large body of work proving similar recovery guarantees for measurement ensembles that feature less randomness [GKK15b; KRT17; KRT17; KZG16] or additional structure tailored to specific applications [CSV13; GKK17; Vor13; Kue15]. The main result of this chapter presented in the next section is of this type.

3.3. Theory

The main theoretical contribution of this chapter is a recovery guarantee for a measurement ensemble motivated by the experimental architecture of linear optical devices: The *randomly erased complex Rademacher* (RECR) ensemble defined below is easier to implement experimentally by the circuit depicted in Fig. 3.1 than e.g. the uniform ensemble. In Section 3.3.1, we introduce this sampling scheme and show that it can be used for phase retrieval. The main technical proof related to the previous section can be found in Section 3.3.2. Section 3.3.3 is concerned with applying the ideas from phase retrieval to the problem of recovering transfer matrices of linear optical circuits. We also discuss the advantages of the RECR ensemble for our particular application in this section.

3.3.1. The RECR ensemble

The uniform ensemble introduced in Section 3.2 is well-suited for theoretical analysis. However, this sampling scheme places high demands on practical implementations as

it necessitates the ability to prepare any input state $|\alpha\rangle$ with α from the complex unit sphere. Therefore, we propose an alternative measurement ensemble that lends itself better to implementations in linear optics: For $p \in [0,1]$, we define a *randomly erased complex Rademacher* (RECR) random variable a to be distributed according to

$$a = \begin{cases} +1 & \text{with prob. } p/4 \\ +\mathrm{i} & \text{with prob. } p/4 \\ 0 & \text{with prob. } 1-p\,. \\ -\mathrm{i} & \text{with prob. } p/4 \\ -1 & \text{with prob. } p/4 \end{cases} \tag{3.15}$$

Here, the constant $1-p$ is referred to as the *erasure probability*. For the (unnormalized) RECR measurement model, we sample the components $\alpha_k^{(l)}$ of the input vectors $\alpha^{(l)}$ according to Eq. (3.15). We also consider the normalized version below. Since there are only four different values for the phases of the components (3.15), the RECR scheme is easier to implement experimentally as we discuss in Section 3.3.3.

Note that a non-zero erasure probability is crucial for phase retrieval as the following example shows: Denote by e_j the canonical basis vectors. For $p = 1$, the complex Rademacher measurement vectors cannot distinguish between the signals $x = e_1$ and $x = e_2$ from measurements (3.9) even in the idealized noiseless case $\epsilon^{(l)} = 0$.

The rest of this section is devoted to proving recovery guarantees for the RECR ensemble for phase retrieval via PhaseLift (3.12). In order to be more self-contained, we also provide recovery guarantees for the Gaussian and uniform ensembles, which are well known [CL14; DH14]. For this purpose, we develop a unified proof strategy inspired by Ref. [DLR16], who derived strong result for sparse vector recovery using similar assumptions, and Ref. [Kab+16] in the non-commutative setting. In the following we consider ensembles of measurement vectors α that satisfy the following moment conditions.

Definition 3.1. *A random vector $\alpha \in \mathbb{C}^n$ is said to be* attentive *if there are positive constants C_{I}, C_{SI}, and C_{SG} such that the following conditions are satisfied.*

- Isotropy on $\mathbb{C}^n$: *for every $z \in \mathbb{C}^n$*

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^2\right] = C_{\mathrm{I}} \|z\|_{\ell_2}^2 \tag{3.16}$$

- Sub-Isotropy on $\mathbb{H}^n$: *denote by $\mathbb{H}^n$ the set of all Hermitian $n \times n$ matrices. Then, we the following condition should hold for every $Z \in \mathbb{H}^n$*

$$\mathbb{E}\left[\langle \alpha, Z\alpha\rangle^2\right] \geq C_{\mathrm{SI}} \|Z\|_2^2 \tag{3.17}$$

- Sub-Gaussian tail behavior: *For every normalized* $z \in \mathbb{C}^n$ *(*$\|z\|_{\ell_2} = 1$*),* $|\langle \alpha, z\rangle|$ *is sub-Gaussian in the sense that its moments obey*

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^{2N}\right] \leq C_{\mathrm{SG}} N! \quad N \in \mathbb{N}. \tag{3.18}$$

The following proposition shows that these conditions are satisfied by the unnormalized Gaussian and RECR ensembles as well as the normalized uniform ensemble. However, we do not have a proof that the practically relevant normalized RECR ensemble is attentive. Therefore, we are going to treat it separately below.

Proposition 3.2. *The following measurement ensembles are attentive according to Definition 3.1:*

1. Gaussian sampling scheme: $\alpha \in \mathbb{C}^n$ *chosen from the standard complex normal distribution* $\mathcal{N}(0, \frac{1}{2}\mathbb{1}_n) + \mathrm{i}\mathcal{N}(0, \frac{1}{2}\mathbb{1}_n)$*. In this case*

$$C_{\mathrm{I}} = C_{\mathrm{SI}} = C_{\mathrm{SG}} = 1. \tag{3.19}$$

2. Uniform sampling scheme: $\alpha \in \mathbb{C}^n$ *chosen uniformly from the complex sphere with radius* $\sqrt{n}$*. In this case*

$$C_{\mathrm{I}} = 1,\ C_{\mathrm{SI}} = \frac{n}{n+1},\ C_{\mathrm{SG}} = \prod_{k=1}^{N-1} \frac{n}{n+k} \leq 1. \tag{3.20}$$

3. (unnormalized) Randomly Erased Complex Rademacher (RECR) sampling scheme: *the components of* $\alpha \in \mathbb{C}^n$ *are chosen independently from the distribution* (3.15)*. The constants depend only on the erasure probability* $1 - p \in [0, 1]$*:*

$$C_{\mathrm{I}} = p,\ C_{\mathrm{SI}} = p \min\{p, 1-p\},\ C_{\mathrm{SG}} = \mathrm{e}^{\frac{3}{2}}. \tag{3.21}$$

Proof. For case 1, consider $\alpha \in \mathbb{C}^n$ be a standard (complex) Gaussian vector and fix any $z \in \mathbb{C}^n$. Then, the random variable $\langle \alpha, z\rangle$ is an instance of a standard (complex normal) random variable $a = \frac{\|z\|_{\ell_2}}{\sqrt{2}}(a_R + i a_I)$ with $a_R, a_I \sim \mathcal{N}(0,1)$. In turn, $|a|^2 = \frac{\|z\|_{\ell_2}^2}{2}(a_R^2 + a_I^2)$ is a rescaled version of a χ^2-distributed random variable with two degrees of freedom. The moments of such a random variable are well-known and we obtain

$$\mathbb{E}(|\langle \alpha, z\rangle|^{2N}) = \left(\frac{\|z\|_{\ell_2}}{\sqrt{2}}\right)^N \times 2^N N! = \|z\|_{\ell_2}^N N!\,. \tag{3.22}$$

From this, we can readily infer $C_{\mathrm{SG}} = 1$, and the special case $N = 1$ yields $C_{\mathrm{I}} = 1$.

For the remaining expression, use an eigenvalue decomposition $Z = \sum_{k=1}^{d} \zeta_k |z^{(k)}\rangle\langle z^{(k)}|$ (with normalized eigenvectors $z^{(k)} \in \mathbb{C}^n$) and note that the random variables $|\langle a, z^{(1)}\rangle|, \ldots,$ $|\langle a, z^{(n)}\rangle|$ are independently distributed and obey Eq. (3.22). Consequently:

$$\mathbb{E}\left[\mathrm{tr}\,(AZ)^2\right] = \mathbb{E}\left[\left(\sum_{k=1}^{d} \zeta_k |\langle \alpha, z^{(k)}\rangle|^2\right)^2\right] \tag{3.23}$$

$$= \sum_{k\neq l} \zeta_k \zeta_l \mathbb{E}\left[|\langle \alpha, z^{(k)}\rangle|^2\right] \mathbb{E}\left[|\langle a, z^{(l)}\rangle|^2\right] + \sum_{k=1}^{d} \zeta_k^2 \mathbb{E}\left[|\langle a, z^{(k)}\rangle|^4\right] \tag{3.24}$$

$$= \sum_{k\neq l} \zeta_k \zeta_l \|z^{(k)}\|_{\ell_2}^2 \|z^{(l)}\|_{\ell_2}^2 + 2\sum_{k=1}^{d} \zeta_k^2 \|z^{(k)}\|_{\ell_2}^4 = \sum_{k,l=1}^{d} \zeta_k \zeta_l + 2\sum_{k=1}^{d} \zeta_k^2 \tag{3.25}$$

$$= \mathrm{tr}(Z)^2 + \mathrm{tr}(Z^2) \geq \|Z\|_2^2, \tag{3.26}$$

which implies $C_{\mathrm{SI}} = 1$.

Now consider the case 2, where α is chosen uniformly from the complex sphere with radius $\sqrt{n}$. This in turn implies that the distribution of $\alpha \in \mathbb{C}^n$ is invariant under arbitrary unitary transformations. Techniques from representation theory – more precisely: Schur's Lemma – then imply

$$\mathbb{E}\left[(|\alpha\rangle\langle\alpha|)^{\otimes N}\right] = n^N \binom{n+N-1}{N}^{-1} P_{\bigvee^N}, \tag{3.27}$$

see e.g. [Sco06, Lemma 1]. Here, $P_{\bigvee^N}$, denotes the projector onto the totally symmetric subspace $\bigvee^N \subseteq (\mathbb{C}^n)^{\otimes N}$. Note that $(|z\rangle\langle z|)^{\otimes N} \in \bigvee^N$ and, moreover $2\mathrm{tr}\left(P_{\vee^2} Z^2\right) = \|Z\|_2^2 + \mathrm{tr}(Z)^2$ for any matrix Z, see e.g. [KZG16, Lemma 17]. Consequently,

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^2\right] = \mathrm{tr}\left(|z\rangle\langle z| \,\mathbb{E}\left[|\alpha\rangle\langle\alpha|\right]\right) = \mathrm{tr}\left(|z\rangle\langle z|\mathbb{I}\right) = \|z\|_{\ell_2}^2, \tag{3.28}$$

$$\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right] = \mathrm{tr}\left(\mathbb{E}\left[(|\alpha\rangle\langle\alpha|)^{\otimes 2}\right] Z^{\otimes 2}\right) = \frac{n}{n+1}\left(\|Z\|_2^2 + \mathrm{tr}(Z)^2\right) \geq \frac{n}{n+1}\|Z\|_2^2, \tag{3.29}$$

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^{2N}\right] = \mathrm{tr}\left(\mathbb{E}\left[(|\alpha\rangle\langle\alpha|)^{\otimes N}\right] (|z\rangle\langle z|)^{\otimes N}\right) = n^N \binom{n+N-1}{N}^{-1} \|z\|_{\ell_2}^{2N} \tag{3.30}$$

$$= N! \frac{n^N (n-1)!}{(n+N-1)!} \leq N!, \tag{3.31}$$

which implies $C_{\mathrm{I}} = 1$, $C_{\mathrm{SI}} = \frac{n}{n+1}$ and $C_{\mathrm{SG}} = 1$.

Finally, consider the case 3, with α sampled from the unnormalized RECR ensemble. Let $\alpha_k = \langle e_k, \alpha \rangle$, where $e_1, \ldots, e_n$ is the orthonormal basis with respect to which the RECR vector is defined. Theses components obey $\mathbb{E}[\alpha_k] = \mathbb{E}[\alpha_k^*] = 0$, as well as $\mathbb{E}\left[|\alpha_k|^2\right] = p$. For any $z \in \mathbb{C}^n$ we then have

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^2\right] = \sum_{i,j=1}^{n} \mathbb{E}\left[\alpha_i^* \alpha_j\right] \langle e_i|z\rangle\langle z|e_j\rangle = p \sum_{i=1}^{n} |\langle e_i, z\rangle|^2 = p\|z\|_{\ell_2}^2. \tag{3.32}$$

Now, Fix $Z \in \mathbb{H}^n$ and compute

$$\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right] = \sum_{i,j,k,l} \mathbb{E}\left[\bar{\alpha}_i \alpha_j {\alpha'_k}^* \alpha'_l\right] \langle e_i|Z|e_j\rangle\langle e_k|Z|e_l\rangle \tag{3.33}$$

$$= \sum_i \mathbb{E}\left[|\alpha_i|^4\right] \langle e_i|Z|e_i\rangle^2 + \sum_{i\neq k} \mathbb{E}\left[|\alpha_i|^2|\alpha_k|^2\right] \left(\langle e_i|Z|e_i\rangle\langle e_k|Z|e_k\rangle + \langle e_i|Z|e_k\rangle\langle e_k|Z|e_i\rangle\right) \tag{3.34}$$

$$= p\sum_{i=1}^{n}\langle e_i|Z|e_i\rangle^2 + p^2 \sum_{i\neq k} \left(\langle e_i|Z|e_i\rangle\langle e_k|Z|e_k\rangle + \langle e_i|Z|e_k\rangle\langle e_k|Z|e_i\rangle\right) \tag{3.35}$$

$$= p^2 \sum_{i,k=1}^{n} \left(\langle e_i|Z|e_i\rangle\langle e_k|Z|e_k\rangle + \langle e_i|Z|e_k\rangle\langle e_k|Z|e_i\rangle\right) + p(1-2p)\sum_{i=1}^{n}\langle e_i|Z|e_i\rangle^2 \tag{3.36}$$

$$= p^2\left(\mathrm{tr}(Z)^2 + \|Z\|_2^2\right) + p(1-2p)\sum_{i=1}^{n}\langle e_i|Z|e_i\rangle^2 \tag{3.37}$$

$$\geq p^2\|Z\|_2^2 + p(1-p)\sum_{i=1}^{n}\langle e_i|Z|e_i\rangle^2 \tag{3.38}$$

To proceed, we consider the following two cases:

$p \leq 1/2$: This implies $p(1-2p) \geq 0$ and consequently

$$\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right] \geq p^2\|Z\|_2^2. \tag{3.39}$$

$p \geq 1/2$: Use $\sum_{i=1}^{n}\langle i|X|i\rangle^2 \leq \|X\|_2^2$ to conclude

$$\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right] \geq (p^2 - p|1-2p|)\|Z\|_2^2 = p(1-p)\|Z\|_2^2. \tag{3.40}$$

This shows that the RECR example is sub-isotropic on $\mathbb{H}^n$.

Finally, fix $z \in \mathbb{C}^n$ with $\|z\|_{\ell_2} = 1$ and note that $|\alpha_k| \leq 1$ together with the independence of α_k, α_l for $k \neq l$ implies

$$\begin{aligned}\mathbb{E}\left[\exp\left(|\langle \alpha, z\rangle|^2\right)\right] =& \mathbb{E}\left[\prod_{k=1}^{n} \exp\left(|\alpha_k|^2 |z_k|^2\right) \prod_{k\neq l} \exp\left(\alpha_k^* \alpha_l z_k^* z_l\right)\right] \\ &\leq \exp\left(\|z\|_{\ell_2}^2\right) \prod_{k\neq l} \mathbb{E}\left[\exp\left(\alpha_k^* \alpha_l z_k^* z_l\right)\right]. \end{aligned} \tag{3.41}$$

Now note that for $k \neq l$, $\alpha_k^* \alpha_l$ is again a RECR random variable $\tilde{\alpha}_{k,l}$, but with erasure probability $1-p^2$. Moreover, every RECR random variable α can be decomposed into the product of two independent random variables: $\alpha = \eta\omega$, where η is a Rademacher random variable and $\omega \in \{0, 1, i\}$ obeys $|\omega| \leq 1$. Consequently

$$\mathbb{E}\left[\exp\left(\bar{\alpha}_k \alpha_l \bar{z}_k z_l\right)\right] = \mathbb{E}\left[\exp\left(\tilde{\alpha}_{k,l} \bar{z}_k z_l\right)\right] = \mathbb{E}_\omega\left[\mathbb{E}_\eta\left[\eta\omega \bar{z}_k z_l\right]\right] = \mathbb{E}_\omega\left[\cos\left(\omega \bar{z}_k z_l\right)\right] \tag{3.42}$$

$$\leq \mathbb{E}_\omega\left[\exp\left(|\omega \bar{z}_k z_l|^2/2\right)\right] \leq \exp\left(\frac{|z_k|^2 |z_l|^2}{2}\right), \tag{3.43}$$

where we have used the standard estimate $\cosh(x) \leq \exp\left(|x|^2/2\right)\ \forall x \in \mathbb{C}$, as well as $|\omega| \leq 1$. Inserting this bound into (3.41) yields

$$\mathbb{E}\left[\exp\left(|\langle \alpha, z\rangle|^2\right)\right] \leq \exp\left(\|z\|_2^2\right) \prod_{k\neq l} \exp\left(\frac{|z_k|^2 |z_l|^2}{2}\right) \leq \exp\left(\|z\|_2^2 + \frac{1}{2}\|z\|_{\ell_2}^4\right) = \mathrm{e}^{\frac{3}{2}}, \tag{3.44}$$

because $\|z\|_{\ell_2} = 1$. Markov's inequality shows that this exponential bound implies a subexponential tail bound for the random variable $|\langle \alpha, z\rangle|^2$:

$$\mathbb{P}\left[|\langle \alpha, z\rangle|^2 \geq t\right] = \mathbb{P}\left[\exp\left(|\langle \alpha, z\rangle|^2\right) \geq \exp(t)\right] \leq \frac{\mathbb{E}\left[\exp\left(|\langle \alpha, z\rangle|^2\right)\right]}{\exp(t)} \leq \mathrm{e}^{\frac{3}{2}-t}. \tag{3.45}$$

This in turn implies the following bound on the moments:

$$\mathbb{E}\left[|\langle \alpha, z\rangle|^{2N}\right] = N \int_0^\infty \mathbb{P}\left[|\langle \alpha, z\rangle|^2 \geq t\right] t^{N-1} \mathrm{d}t \leq N\mathrm{e}^{\frac{3}{2}} \int_0^\infty \mathrm{e}^{-t} t^{N-1} \mathrm{d}t = \mathrm{e}^{\frac{3}{2}} N!, \tag{3.46}$$

where we have used a well-known integration formula for moments, see e.g. [FR13, Prop. 7.1], as well as integration by parts. □

The conditions in Definition 3.1 naturally appear in the proof of the following theorem, which is the fundamental technical part of this work. It is used to provide rigorous recovery guarantees for phase retrieval via PhaseLift.

Proposition 3.3. *Suppose that $m = Cn$ vectors $\alpha^{(1)}, \dots, \alpha^{(m)} \in \mathbb{C}^n$ have been chosen independently at random from an attentive ensemble. Let $Z \geq 0$ and $x \in \mathbb{C}^n$. Then, the measurement operator*

$$\mathcal{A}(Z) = \sum_l \operatorname{tr}\left(|\alpha^{(l)}\rangle\langle\alpha^{(l)}|Z\right) e_l \tag{3.47}$$

satisfies

$$\|Z - |x\rangle\langle x|\|_2 \leq \frac{1}{m} \max\left\{\tau, \frac{6}{\nu}\right\} \|\mathcal{A}\left(Z - |x\rangle\langle x|\right)\|_{\ell_1}. \tag{3.48}$$

with probability at least $1 - 3\mathrm{e}^{-\gamma m}$. Here, C and γ denote suitable positive constants.

Note that the measurement operator notation (3.47) is simply a shorthand for

$$y^{(l)} = \langle \alpha^{(l)}, Z\alpha^{(l)} \rangle \qquad (l = 1, \dots, m). \tag{3.49}$$

We postpone the proof of this proposition to Section 3.3.2. Instead, we consider a slight generalization of the properties in Definition 3.1, which are the most general class of measurement ensembles we consider in this work.

Definition 3.4. *We say that a random vector $\alpha \in \mathbb{C}^n$ is* super-attentive*, if there is an attentive random vector $\tilde{\alpha}$ and a function $f\mathbb{C}^n \to \mathbb{R}$ with $f(\tilde{\alpha}) \geq 1$ almost surely such that $\alpha = f(\tilde{\alpha})\colon \tilde{\alpha}$.*

The main example of a super-attentive distribution in this work is a normalized RECR vector with length $\sqrt{n}$. Denote by $\tilde{\alpha}$ an unnormalized RECR vector and set $f(\tilde{\alpha}) = \sqrt{n}/\|\tilde{\alpha}\|_{\ell_2}$, then

$$\alpha = f(\tilde{\alpha})\tilde{\alpha} \tag{3.50}$$

is a normalized RECR vector. Note that every attentive random vector is super-attentive trivially.

Let us now state the central result of this section, namely a recovery guarantee for super-attentive measurement ensembles. The following theorem is a substantial generalization of existing results regarding Gaussian and uniform measurement ensembles [CL14; DH14].

Theorem 3.5. *Suppose that $m = Cn$ vectors $\alpha^{(1)}, \dots, \alpha^{(m)} \in \mathbb{C}^n$ have been chosen independently at random from a super-attentive ensemble. Then, the optimizer $X^\sharp$ of the convex program* (3.12) *satisfies*

$$\left\|X^\sharp - |x\rangle\langle x|\right\|_2 \leq \frac{C'\|\epsilon\|_{\ell_1}}{m}. \tag{3.51}$$

with probability at least $1 - 3\mathrm{e}^{-\gamma m}$ *for some constant* $\gamma > 0$. *Here,* $\|\cdot\|_2$ *denotes the Hilbert-Schmitt norm* $\|Z\|_2^2 = \mathrm{tr}\left(ZZ^\dagger\right)$, *while* C, C' *and* γ *represent constants of sufficient size. Furthermore,* $\|\epsilon\|_{\ell_1}$ *is a bound on the total noise of all measurements* (3.9)

$$\|\epsilon\|_{\ell_1} = \sum_{l=1}^{m} |\epsilon^{(l)}|. \tag{3.52}$$

Proof of Theorem 3.5. Denote by $\tilde{\alpha}^{(l)}$ the attentive vector corresponding to $\alpha^{(l)}$ and by f the scaling function from Definition 3.4. Then, the measurements outcomes $y^{(l)}$ can be mapped to measurement outcomes of $\tilde{\alpha}^{(l)}$ by

$$\tilde{y}^{(l)} := \frac{1}{f(\tilde{\alpha})^2}\, y^{(l)} = \left|\langle \tilde{\alpha}^{(l)}, x\rangle\right|^2 + \tilde{\epsilon}^{(l)}. \tag{3.53}$$

with the rescaled error vectors given by

$$\tilde{\epsilon}^{(l)} = \frac{1}{f(\tilde{\alpha})^2}\, \epsilon^{(l)}. \tag{3.54}$$

Now, Proposition 3.3 implies that a measurement operator $\tilde{\mathcal{A}}$ containing $m \geq Cn$ measurements sampled from an attentive distribution satisfies Eq. (3.48) with probability at least $1 - 3\mathrm{e}^{-\gamma m}$. Conditioned on this event, we have for any $Z \geq 0$ and $x \in \mathbb{C}^n$

$$\|Z - |x\rangle\langle x|\|_2 \leq \frac{C'}{2m}\left\|\tilde{\mathcal{A}}(Z) - \tilde{y} + \tilde{\epsilon}\right\|_{\ell_1} \leq \frac{C'}{2m}\left(\|\tilde{\epsilon}\|_{\ell_1} + \|\tilde{\mathcal{A}}(Z) - \tilde{y}\|\right), \tag{3.55}$$

with $C' = 2\max\{\tau, 6/\nu\}$. For the first summand, we have.

$$\|\tilde{\epsilon}\|_{\ell_1} = \sum_l \left|\frac{1}{f(\tilde{\alpha}^{(l)})^2}\, \epsilon^{(l)}\right| \leq \|\epsilon\|_{\ell_1} \tag{3.56}$$

since $f(\tilde{\alpha}^{(l)}) \geq 1$. For the second summand on the right hand side of Equation (3.55), the same argument gives

$$\|\tilde{\mathcal{A}}(Z) - \tilde{y}\|_{\ell_1} = \sum_l \left|\langle \tilde{\alpha}^{(l)}, Z\tilde{\alpha}^{(l)}\rangle - \tilde{y}^{(l)}\right| \tag{3.57}$$

$$= \sum_l \frac{1}{f(\tilde{\alpha}^{(l)})^2}\left|\langle \alpha^{(l)}, Z\alpha^{(l)}\rangle - y^{(l)}\right| \tag{3.58}$$

$$\leq \|\mathcal{A}(Z) - y\|_{\ell_1} \tag{3.59}$$

PhaseLift – the convex optimization problem (3.12) – minimizes the right hand side of this bound over all $Z \geq 0$. Since $Z = |x\rangle\langle x|$ is a feasible point of this optimization, we can conclude that the minimizer $Z^\sharp$ obeys

$$\|\mathcal{A}(Z^\sharp) - y\|_{\ell_1} \leq \|\mathcal{A}(|x\rangle\langle x|) - y\|_{\ell_1} = \|\epsilon\|_{\ell_1} \tag{3.60}$$

which concludes the proof. □

The constants C, C', and γ implicitly depend on the ensemble constants C_{I}, C_{SI}, and C_{SG} and can in principle be extracted from the proof. Note that although we are recovering $|x\rangle\langle x|$, which is embedded in the n^2 dimensional space of all $n \times n$ matrices, the demand on the number of measurements m in Theorem 3.5 scales linearly in the original problem's dimension n. This is optimal up to the constant multiplicative factor C. Analytical bounds on this constant C are usually too pessimistic to be practical and it is widely believed that $m = 4n - 4$ such measurements are actually sufficient, that is $C = 4 + o(n)$ [HMW13].

Recall from Eq. (3.14) that we obtain the recovery of the signal vector $x^\sharp$ from the minimizer $X^\sharp$ of the PhaseLift program Eq. (3.12) via an eigenvalue decomposition. In [CL14] it was shown that Eq. (3.51) implies

$$\min_{0\leq\phi\leq 2\pi} \|x^\sharp - \mathrm{e}^{\mathrm{i}\phi}x\|_{\ell_2} \leq C'' \frac{\|\epsilon\|_{\ell_1}}{m\|x\|_{\ell_2}}, \tag{3.61}$$

where C'' denotes another constant of sufficient size. In words, we are able to recover the original signal x up to a global phase and up to an error that is determined by the signal-to-noise ratio.

3.3.2. Proof of Proposition 3.3

In this section we present a proof for the fundamental Proposition 3.3. Our analysis is inspired by Ref. [DLR16] (who derived strong results for sparse vector recovery using similar assumptions) and Ref. [Kab+16] in the non-commutative setting. Moreover, Krahmer and Liu considered a real-valued version of the problem addressed here [KL18].

Our analysis is based on two fundamental results in random matrix theory. First, the assumption of subgaussian tails (3.18) implies strong bounds on the operator norm of matrices of the form $\sum_{k=1}^m |\alpha^{(l)}\rangle\langle\alpha^{(l)}|$:

Theorem 3.6 (Variant of Theorem 5.35 in [Ver10])**.** *Suppose that $\alpha^{(1)}, \ldots, \alpha^{(m)}$ are independent copies of a subgaussian random vector obeying Eq.* (3.18) *with constant C_{SG}. Let*

$$\tilde{H} = \frac{1}{m}\sum_{k=1}^{m} \left(a_k |\alpha^{(l)}\rangle\langle\alpha^{(l)}| - \mathbb{E}\left[a_k |\alpha^{(l)}\rangle\langle\alpha^{(l)}| \right] \right), \tag{3.62}$$

where $a_k \in \mathbb{C}$ and $|a_k| \leq 1$. Then,

$$\mathbb{P}\left[\|\tilde{H}\|_{2\to 2} \geq t \right] \leq \begin{cases} 2\exp\left(2\ln(3)n - \frac{mt^2}{8C_{\mathrm{SG}}} \right) & 0 \leq t \leq 2C_{\mathrm{SG}}, \\ 2\exp\left(2\ln(3)n - \frac{m}{2}\left(t - C_{\mathrm{SG}}\right) \right) & t \geq 2C_{\mathrm{SG}}, \end{cases} \tag{3.63}$$

where $\|\cdot\|_{2\to 2}$ denotes the operator norm.

The second result is a generalization of "Gordon's escape through a mesh"-Theorem [Gor88] (a random subspace avoids a subset provided the subset is small in some sense). The version we use here is due to Mendelson [Men14; KM15], see also see also [Tro15].

Theorem 3.7 (Mendelson's small ball method)**.** *Suppose that the measurement operator $\mathcal{A} : \mathbb{H}^n \to \mathbb{R}^m$ contains m independent copies $A^{(l)}$ of a random matrix $A \in \mathbb{H}^n$, that is*

$$\mathcal{A}(Z) = \sum_{l=1}^{m} \operatorname{tr}(A^{(l)} Z)\, e_l \tag{3.64}$$

with e_l denoting the l-th canonical basis vector. For $D \subseteq \mathbb{H}^n$ and $\xi > 0$ define

$$Q_\xi(D, A) = \inf_{Z \in D} \mathbb{P}\left[|\operatorname{tr}(A^{(l)} Z)| \geq \xi \right] \qquad \textit{(marginal tail funtion)}, \tag{3.65}$$

$$W_m(D, A) = 2\,\mathbb{E}\left[\sup_{Z \in D} \operatorname{tr}(ZH) \right] \qquad \textit{(mean empirical width)}, \tag{3.66}$$

where

$$H = \frac{1}{\sqrt{m}} \sum_{l=1}^{m} \eta_l A^{(l)}. \tag{3.67}$$

Here, the η_l are independent Rademacher random variables, i.e. $\mathbb{P}(\eta_l = 1) = \mathbb{P}(\eta_l = -1) = \frac{1}{2}$. Then for any $\xi > 0$ and $t > 0$

$$\frac{1}{\sqrt{m}} \inf_{Z \in D} \|\mathcal{A}(Z)\|_{\ell_1} \geq \xi \sqrt{m} Q_{2\xi}(D, A) - W_m(D, A) - \xi t \tag{3.68}$$

with probability at least $1 - \mathrm{e}^{-2t^2}$.

The following two propositions summarize several results presented in [Kab+16] and adapt them to the problem of phase retrieval.

Proposition 3.8. *Let $\mathcal{S}^{n^2-1} = \{Z \in \mathbb{H}^n \colon \|Z\|_2 = 1\}$ be the (Frobenius norm) unit sphere and $\mathcal{B}_1 = \operatorname{conv}\left\{ \pm |x\rangle\langle x| \colon x \in \mathcal{S}^{n-1} \right\}$ the trace-norm ball in $\mathbb{H}^n$. Define*

$$D := \mathcal{S}^{d^2-1} \cap 3\mathcal{B}_1. \tag{3.69}$$

Also, let $\mathcal{A}(Z) = \sum_{l=1}^{m} \operatorname{tr}(A^{(l)} Z)\, e_l$ be a measurement operator that obeys

$$\frac{\tau}{m} \|\mathcal{A}(Z)\|_{\ell_1} \geq \|Z\|_2 \quad \forall Z \in D \tag{3.70}$$

$$\Big\| \frac{1}{\nu m} \sum_{l=1}^{m} A^{(l)} - \mathbb{I} \Big\|_\infty \leq \frac{1}{6} \tag{3.71}$$

for some $\tau, \nu > 0$. Then, the following relation holds for any $Z \geq 0$ and any $|x\rangle\langle x|$:

$$\|Z - |x\rangle\langle x|\|_2 \leq \frac{1}{m} \max\left\{ \tau, \frac{6}{\nu} \right\} \|\mathcal{A}\left(Z - |x\rangle\langle x|\right)\|_{\ell_1}. \tag{3.72}$$

Proof. In the proof we will frequently use the decomposition $Z = Z_1 + Z_c$ for Z with eigenvalue decomposition $Z = \sum_{k=1}^{n} \lambda_k |z^{(k)}\rangle\langle z^{(k)}|$. Assuming $\lambda_1 \geq \ldots \geq \lambda_n$, $Z_1 = \lambda_1 |z^{(1)}\rangle\langle z^{(1)}|$ is the leading rank-one component and $Z_c = Z - Z_1$ is the "tail". Note that, in particular, $Z = Z_1$ if and only if Z has unit rank. Equation (3.72) is invariant under re-scaling, so we may w.l.o.g. assume $\|Z - |x\rangle\langle x|\|_2 = 1$. We treat the following two cases separately:

$$\text{I.)} \quad \|(Z - |x\rangle\langle x|)_1\|_1 \geq \frac{1}{2}\|(Z - |x\rangle\langle x|)_c\|_1, \tag{3.73}$$

$$\text{II.)} \quad \|(Z - |x\rangle\langle x|)_1\|_1 < \frac{1}{2}\|(Z - |x\rangle\langle x|)_c\|_1. \tag{3.74}$$

Note that I.) implies

$$\begin{aligned} \|Z - |x\rangle\langle x|\|_1 \leq & \|(Z - |x\rangle\langle x|)_1\|_1 + \|(Z - |x\rangle\langle x|)_c\|_1 \leq 3\|(Z - |x\rangle\langle x|)_1\|_1 & (3.75)\\ = & 3\|(Z - |x\rangle\langle x|)_1\|_2 \leq 3\|Z - |x\rangle\langle x|\|_2 = 3 & (3.76) \end{aligned}$$

which in turn implies that $Z - |x\rangle\langle x|$ is contained in $3\mathcal{B}_1$. Thus, (3.70) is applicable and yields

$$\|Z - |x\rangle\langle x|\|_2 \leq \frac{\tau}{m}\|\mathcal{A}(Z - |x\rangle\langle x|)\|_{\ell_1} \tag{3.77}$$

which establishes Eq. (3.72) for the case (3.73).

For the second case, we use a consequence of von Neumann's trace inequality, see e.g. [HJ94, Theorem 7.4.9.1]: Let A, B be matrices with singular values $\sigma_k(A), \sigma_k(B)$ arranged in non-increasing order. Then

$$\|A - B\|_1 \geq \sum_{k=1}^{n} |\sigma_k(A) - \sigma_k(B)| \tag{3.78}$$

This relation implies

$$\begin{aligned} \|Z\|_1 = & \||x\rangle\langle x| - (|x\rangle\langle x| - Z)\|_1 \geq \sum_{k=1}^{n} |\sigma_k(|x\rangle\langle x|) - \sigma_k(|x\rangle\langle x| - Z)| & (3.79)\\ \geq & \sigma_1(|x\rangle\langle x|) - \sigma_1\left(|x\rangle\langle x| - Z\right) + \sum_{k=2}^{n} \sigma_k\left(|x\rangle\langle x| - Z\right) & (3.80)\\ = & \||x\rangle\langle x|\|_1 - \|(|x\rangle\langle x| - Z)_1\|_1 + \|(|x\rangle\langle x| - Z)_c\|_1 & (3.81)\\ > & \||x\rangle\langle x|\|_1 + \frac{1}{2}\|(|x\rangle\langle x| - Z)_c\|_1, & (3.82) \end{aligned}$$

where the last inequality follows from (3.74). Consequently,

$$\begin{aligned} \||x\rangle\langle x| - Z\|_1 = & \|(|x\rangle\langle x| - Z)_1\|_1 + \|(|x\rangle\langle x| - Z)_c\|_1 \leq \frac{3}{2}\|(|x\rangle\langle x| - Z)_c\|_1 \\ < & 3\left(\|Z\|_1 - \||x\rangle\langle x|\|_1\right). & (3.83) \end{aligned}$$

Now, positive semidefiniteness of both Z and $|x\rangle\langle x|$ together with assumption (3.71) implies

$$\begin{aligned}\|Z\|_1 - \||x\rangle\langle x|\|_1 &= \mathrm{tr}(Z - |x\rangle\langle x|) = \mathrm{tr}\left(\mathbb{I}\left(Z - |x\rangle\langle x|\right)\right) && (3.84)\\ &= \mathrm{tr}\left(\left(\mathbb{I} - \frac{1}{\nu m}\sum_{k=1}^{m} A^{(l)}\right) Z - |x\rangle\langle x|\right) + \frac{1}{\nu m}\sum_{k=1}^{m}\mathrm{tr}\left(A_k(Z - |x\rangle\langle x|)\right) && (3.85)\\ &\leq \left\|\mathbb{I} - \frac{1}{\nu m}\sum_{k=1}^{m} A_k\right\|_\infty \|Z - |x\rangle\langle x|\|_1 + \frac{1}{\nu m}\|\mathcal{A}(|x\rangle\langle x| - Z)\|_{\ell_1} && (3.86)\\ &\leq \frac{1}{6}\|Z - |x\rangle\langle x|\|_1 + \frac{1}{\nu m}\|\mathcal{A}(|x\rangle\langle x| - Z)\|_{\ell_1}. && (3.87)\end{aligned}$$

Inserting this into (3.83) yields

$$\||x\rangle\langle x| - Z\|_1 < \frac{1}{2}\||x\rangle\langle x| - Z\|_1 + \frac{3}{\nu m}\|\mathcal{A}(|x\rangle\langle x| - Z)\|_{\ell_1} \tag{3.88}$$

which implies the claim for case II in (3.74). □

Lemma 3.9. *Let D be the set introduced in Eq.* (3.69) *and let $A = |\alpha\rangle\langle\alpha|$, where α satisfies Eqs.* (3.17) *and* (3.18)*. Then, the marginal tail function* (3.65) *obeys*

$$Q_\xi(D, A) \geq C_Q\left(1 - \frac{\xi^2}{C_{\mathrm{SI}}}\right)^2 \quad \forall 0 \leq \xi \leq \sqrt{C_{\mathrm{SI}}}, \tag{3.89}$$

where $C_Q > 0$ is a sufficiently small constant.

Proof. Fix $Z \in D$, then $\|Z\|_2 = 1$ by definition of D. Note that sub-isotropy (3.17) and the Paley-Zygmund inequality imply for any $\xi \in [0, 1]$

$$\begin{aligned}\mathbb{P}\left[|\langle a|Z|a\rangle| \geq \xi\right] \geq& \mathbb{P}\left[\langle\alpha|Z|\alpha\rangle^2 \geq \frac{\xi^2}{C_{\mathrm{SI}}}\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right]\right] && (3.90)\\ &\geq \left(1 - \frac{\xi^2}{C_{\mathrm{SI}}}\right)^2 \frac{\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^2\right]^2}{\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^4\right]}. && (3.91)\end{aligned}$$

Sub-isotropy ensures that the numerator is lower bounded by $C_{\mathrm{SI}}^2\|Z\|_2^4 = C_{\mathrm{SI}}^2$. In order to derive an upper bound on the denominator, we use the constraint $\|Z\|_1 \leq 3$ for any $Z \in D$ together with the subgaussian tail behavior (3.18) of α. Insert an eigenvalue decomposition $Z = \sum_{i=1}^n \lambda_i |z^{(i)}\rangle\langle z^{(i)}|$ (with $\lambda_i \in \mathbb{R}$ and $z^{(i)} \in \mathcal{S}^{n-1}$) and note

$$\mathbb{E}\left[\langle\alpha|Z|\alpha\rangle^4\right] \leq \sum_{i_1,i_2,i_3,i_4=1}^{n} |\lambda_{i_1}\lambda_{i_2}\lambda_{i_3}\lambda_{i_4}|\mathbb{E}\left[\prod_{k=1}^{4}|\langle\alpha, z^{(i_k)}\rangle|^2\right]. \tag{3.92}$$

Now fix $z^{(i_1)}, \dots, z^{(i_4)}$ and use the inequality of arthimetic and geometric means as well as the fundamental relation between ℓ_p-norms ($\|v\|_{\ell_1} \leq k^{1-\frac{1}{k}} \|v\|_{\ell_k}$ for $v \in \mathbb{R}^k$) to conclude

$$\mathbb{E}\left[\prod_{k=1}^{4} |\langle \alpha, z^{(i_k)} \rangle|^2\right] \leq \frac{1}{4} \sum_{k=1}^{4} \mathbb{E}\left[|\langle \alpha, z^{(i_k)} \rangle|^8\right] \leq C_{\mathrm{SG}} 4!, \tag{3.93}$$

where the last inequality follows from condition (3.18). Consequently,

$$\mathbb{E}\left[\langle \alpha | Z | \alpha \rangle^4\right] \leq C_{\mathrm{SG}} 4! \sum_{i_1, i_2, i_3, i_4} |\lambda_{i_1} \lambda_{i_2} \lambda_{i_3} \lambda_{i_4}| = 24 C_{\mathrm{SG}} \|Z\|_1^4 \leq 24 \times 3^4 C_{\mathrm{SG}}, \tag{3.94}$$

because $Z \in D$ implies $\|Z\|_1 \leq 3$. In summary,

$$\mathbb{P}\left[|\langle \alpha | Z | \alpha \rangle| \geq \xi\right] \geq \left(1 - \frac{\xi^2}{C_{\mathrm{SI}}}\right)^2 \frac{\mathbb{E}\left[\langle \alpha | Z | a \rangle^2\right]^2}{\mathbb{E}\left[\langle \alpha | Z | \alpha \rangle^4\right]} \geq \left(1 - \frac{\xi^2}{C_{\mathrm{SI}}}\right)^2 \frac{C_{\mathrm{SI}}^2}{1944 C_{\mathrm{SG}}} \tag{3.95}$$

and the bound on $Q_\xi(D, A)$ with $C_Q = \frac{C_{\mathrm{SI}}^2}{1944 C_{\mathrm{SG}}}$ follows from the fact that this lower bound holds for any $Z \in D$. □

Lemma 3.10 (Bound on the mean empirical width)**.** *Let D be the set introduced in Eq.* (3.69) *and let $H = \frac{1}{\sqrt{m}} \sum_{l=1}^{m} \eta_l |\alpha^{(l)}\rangle\langle\alpha^{(l)}|$, where each $\alpha^{(l)}$ is subexponential in the sense of* (3.18) *and $m \geq \frac{2 \ln(3)}{C_{\mathrm{SG}}} n$. Then there exists a constant $C_{\mathrm{W}} > 0$ such that*

$$W_m(D, A) \leq C_{\mathrm{W}} \sqrt{n}. \tag{3.96}$$

Proof. Note that by construction $D \subseteq 3\mathcal{B}_1$, and consequently,

$$W_m(D, A) = 2\mathbb{E}\left[\sup_{Z \in D} \mathrm{tr}(ZH)\right] \leq 6\mathbb{E}\left[\sup_{Z \in \mathcal{B}_1} \mathrm{tr}(ZH)\right] = 6\mathbb{E}\left[\|H\|_\infty\right], \tag{3.97}$$

where the last equality follows from the duality of trace and operator norm. Now note that $\tilde{H} = \sqrt{m} H$ is of the form (3.62), where each $\alpha^{(l)}$ is an independent Rademacher random variable. Theorem 3.6 thus implies

$$\mathbb{P}\left[\|H\|_\infty \geq t\right] \leq \begin{cases} 2 \times 9^n \exp\left(-\frac{t^2}{8 C_{\mathrm{SG}}}\right) & t \leq 2 C_{\mathrm{SG}} \sqrt{m}, \\ 2 \times 9^n \exp\left(-\frac{\sqrt{m}}{2}\left(t - C_{\mathrm{SG}} \sqrt{m}\right)\right) & t \geq 2 C_{\mathrm{SG}} \sqrt{m} \end{cases} \tag{3.98}$$

and we can bound $\mathbb{E}\left[\|H\|_\infty\right]$ by using the absolute moment formula, see e.g. [FR13, Propostion 7.1], and bounding the effect of the tails via (3.98). To this end, we split

the real line into three intervals $[0, c\sqrt{n}], [c\sqrt{n}, 2C_{\mathrm{SG}}\sqrt{m}], [2C_{\mathrm{SG}}\sqrt{m}, \infty[$, where c is a constant that we fix later:

$$\mathbb{E}\left[\|H\|_\infty\right] = \int_0^\infty \mathbb{P}\left[\|H\|_\infty \geq t\right] \mathrm{d}t \tag{3.99}$$

$$\begin{aligned} &\leq \int_0^{c\sqrt{n}} 1 \mathrm{d}t + 2 \times 9^n \Bigg(\int_{c\sqrt{n}}^{2C_{\mathrm{SG}}\sqrt{m}} 2 \exp\left(-\frac{t^2}{8C_{\mathrm{SG}}}\right) \mathrm{d}t \\ &\qquad + \mathrm{e}^{\frac{mC_{\mathrm{SG}}}{2}} \int_{2C_{\mathrm{SG}}\sqrt{m}}^\infty \exp\left(-\frac{\sqrt{m}t}{2}\right) \mathrm{d}t \Bigg) \end{aligned} \tag{3.100}$$

$$\leq c\sqrt{n} + 2 \times 9^n \left(\int_{c\sqrt{n}}^{2C_{\mathrm{SG}}\sqrt{m}} \exp\left(-\frac{t^2}{8C_{\mathrm{SG}}}\right) \mathrm{d}t + \frac{2}{\sqrt{m}} \mathrm{e}^{-\frac{C_{\mathrm{SG}}m}{2}} \right). \tag{3.101}$$

For the remaining Gauss integral, we use $\frac{t}{c\sqrt{n}} \geq 1 \; \forall t \geq c\sqrt{n}$ to conclude

$$\int_{c\sqrt{n}}^{2C_{\mathrm{SG}}\sqrt{m}} \exp\left(-\frac{t^2}{8C_{\mathrm{SG}}}\right) \mathrm{d}t \leq \int_{c\sqrt{n}}^\infty \frac{t}{c\sqrt{n}} \exp\left(-\frac{t^2}{8C_{\mathrm{SG}}}\right) \mathrm{d}t = \frac{8C_{\mathrm{SG}}}{c\sqrt{n}} \exp\left(-\frac{c^2 n}{8C_{\mathrm{SG}}}\right). \tag{3.102}$$

Now, fixing $c = 4\sqrt{\ln(3) C_{\mathrm{SG}}}$ assures $\exp\left(-\frac{c^2 n}{8C_{\mathrm{SG}}}\right) = 9^{-n}$ and consequently

$$\mathbb{E}\left[\|H\|_\infty\right] \leq 4\sqrt{\ln(3) C_{\mathrm{SG}} n} + \frac{4\sqrt{C_{\mathrm{SG}}}}{\sqrt{\ln(3)n}} + \frac{4}{\sqrt{m}} \mathrm{e}^{2\ln(3)n - C_{\mathrm{SG}}m} \tag{3.103}$$

$$\leq 4\sqrt{C_{\mathrm{SG}}} \left(\sqrt{\ln(3)n} + \frac{2}{\sqrt{\ln(3)n}} \right) \leq 12\sqrt{\ln(3) C_{\mathrm{SG}} n}. \tag{3.104}$$

where the second inequality follows from $m \geq \frac{2\ln(3)}{C_{\mathrm{SG}}} n$. Inserting this bound into (3.97) yields the claim with $C_{\mathrm{W}} = 72\sqrt{\ln(3) C_{\mathrm{SG}}}$. □

Proof of Proposition 3.3. Now we are ready to apply Mendelson's small ball method (3.68). For D defined in (3.69) and measurements $A_l = |\alpha^{(l)}\rangle\langle\alpha^{(l)}|$ with α_l obeying Eqs. (3.17) and (3.18), the bounds from the previous Lemmas imply

$$\frac{1}{\sqrt{m}} \inf_{Z \in D} \|\mathcal{A}(Z)\|_{\ell_1} \geq \xi\sqrt{m} C_Q \left(1 - \frac{4\xi^2}{C_{\mathrm{SI}}}\right)^2 - 2C_{\mathrm{W}}\sqrt{n} - \xi t \quad \forall \xi \in (0, 1/\sqrt{C_{\mathrm{SI}}}), \forall t \geq 0 \tag{3.105}$$

with probability at least $1 - \mathrm{e}^{-2t^2}$. We choose $\xi = \sqrt{C_{\mathrm{SI}}}/4$ and $t = \gamma_1\sqrt{m}$, where

$\gamma_1 = \frac{9C_Q}{32}$ and obtain with probability at least $1 - \exp(-2\gamma_1 m)$:

$$\frac{1}{\sqrt{m}} \inf_{Z \in D} \|\mathcal{A}(Z)\|_{\ell_1} \geq \frac{9C_Q\sqrt{C_{\mathrm{SI}}}}{64}\sqrt{m} - C_{\mathrm{W}}\sqrt{n} - \frac{\sqrt{C_{\mathrm{SI}}}}{4}\frac{9C_Q}{32}\sqrt{m} \tag{3.106}$$

$$= C_{\mathrm{W}}\left(\frac{9C_Q\sqrt{C_{\mathrm{SI}}}}{128C_{\mathrm{W}}}\sqrt{m} - \sqrt{n}\right). \tag{3.107}$$

Setting $m = Cn$ with $C = \left(\frac{256C_{\mathrm{W}}}{9C_Q\sqrt{C_l}}\right)^2$ implies

$$\frac{1}{\sqrt{m}} \inf_{Z \in D} \|\mathcal{A}(Z)\|_{\ell_1} \geq 2C_{\mathrm{W}}\sqrt{n} = \frac{2C_{\mathrm{W}}}{\sqrt{C}}\sqrt{m} \tag{3.108}$$

with probability at least $1 - \mathrm{e}^{-2\gamma_1 m}$. For $\tau = \frac{2C_{\mathrm{W}}}{\sqrt{C}}$, the first claim in Proposition 3.3 follows from rearranging this expression and using $\|Z\|_2 = 1$ for all $Z \in D$.

Let us now move on to establishing the second statement (3.71): Isotropy (3.16) implies

$$\frac{1}{C_{\mathrm{I}}m}\sum_{l=1}^{m} |\alpha^{(l)}\rangle\langle\alpha^{(l)}| - \mathbb{I} = \frac{1}{C_{\mathrm{SG}}m}\sum_{l=1}^{m}\left(|\alpha^{(l)}\rangle\langle\alpha^{(l)}| - \mathbb{E}\left[|\alpha^{(l)}\rangle\langle\alpha^{(l)}|\right]\right) \tag{3.109}$$

and each $\alpha^{(l)}$ has subgaussian tails by assumption (3.18). Thus, Theorem 3.6 is applicable and setting $t = \min\left\{\frac{1}{6}, 2C_{\mathrm{SG}}\right\}$ yields

$$\mathbb{P}\left[\left\|\frac{1}{C_{\mathrm{I}}m}\sum_{l=1}^{m} |\alpha^{(l)}\rangle\langle\alpha^{(l)}| - \mathbb{I}\right\|_\infty \geq \frac{1}{6}\right] \leq 2\exp\left(2\ln(3)n - \frac{C_{\mathrm{I}}m\min\{1/6, 2C_{\mathrm{SG}}\}}{8C_{\mathrm{SG}}}\right) \tag{3.110}$$

$$\leq 2\exp(-\gamma_2 m), \tag{3.111}$$

where the second inequality follows from $m \geq Cn$, provided that C is sufficiently large. Finally, we use the union bound for the overall probability of failure and set $\gamma := \min\{2\gamma_1, \gamma_2\}$. □

3.3.3. Characterization via PhaseLift

In this section, we are going to apply the results from the last section to the original problem of recovering the transfer matrix of a linear optical circuit. The measured intensity at detector j as given by Eq. (3.4) exclusively provides us with information about the j-th row vector of M:

$$I_j(\alpha) = \left|\sum_{k=1}^{n} M_{j,k}\alpha_k\right|^2 + \epsilon_j = |\langle M_j{}^*, \alpha\rangle|^2 + \epsilon_j. \quad 1 \leq j \leq n \tag{3.112}$$

Here, we have defined M_j as the row vectors of M. Since the measured intensities in Eq. (3.112) exactly resemble the measurement model of the phase retrieval problem in Eq. (3.9), we can use the ideas introduced in Section 3.2 to recover M: For this purpose, we propose the following protocol:

1. sample m random coherent input vectors $\alpha^{(l)}$ from an appropriate ensemble,
2. measure the $m \times n$ intensities $I_1(\alpha^{(l)}), \ldots, I_n(\alpha^{(l)})$ with $l = 1, \ldots, m$, and
3. use PhaseLift (3.12) to recover each M_j individually.

In Section 3.3.1, we introduced multiple ensembles to choose the $\alpha^{(l)}$ from. However, not all of these are equally well suited for the problem at hand. First and foremost, we recall from Section 3.1 that input vectors with constant norm $\|\alpha^{(l)}\|_{\ell_2} = 1$ are better suited for the setup depicted in Fig. 3.1. Also, recall that the RECR sampling scheme was conceived with our application in linear optics in mind: One major drawback of the uniform scheme is that each component may take any possible value for its complex phase. In contrast, the RECR scheme has only four possible values for the phase shift, namely $\frac{k\pi}{2}$ for $k = 1, \ldots, 4$. Therefore, the reconfigurable phase shifters in the implementation outlined in Fig. 3.1 can be calibrated to these values. A similar argument applies to the magnitudes of the RECR components, which can only assume n possible values due to the additional normalization constraint $\|\alpha^{(l)}\|_{\ell_2} = 1$. However, the current linear architecture does not benefit from this additional constraints. Using a tree-like structure in the preparation stage could further improve the practical performance of the PhaseLift reconstruction using RECR vectors. We discuss this idea further in the conclusion and outlook section.

The rest of this section is devoted to adapting the results from Section 3.3.1 to derive rigorous performance guarantees for the proposed characterization protocol outlined above. We start by stating its rigorous version.

Protocol 3.11 (*Reconstruction of the transfer matrix M*)**.** *Let M be an arbitrary $n \times n$ transfer matrix as defined in* (3.3)*. In order to approximately recover it, sample $m = Cn$ random coherent input states $|\alpha^{(1)}\rangle, \ldots, |\alpha^{(m)}\rangle$, with $\alpha^{(l)}$ chosen from the uniform or RECR scheme normalized such that $\|\alpha^{(l)}\|_{\ell_2} = 1$. Measure the mn intensities*

$$y_j^{(l)} = \left| \sum_i M_{j,i}\, \alpha_i^{(l)} \right|^2 + \epsilon_j^{(l)} \quad \forall 1 \leq j \leq n, \quad 1 \leq l \leq m, \tag{3.113}$$

where $\epsilon_j^{(l)}$ denotes the additive noise at detector site j when measuring the intensity

resulting from input state $|\alpha^{(l)}\rangle$*. For each* $1 \leq j \leq n$*, solve the semi-definite program*

$$\begin{aligned} Z_j^\sharp = \underset{Z \in \mathbb{H}^n}{\operatorname{argmin}} \quad & \sum_{l=1}^{m} \left| \operatorname{tr}\left((|\alpha^l\rangle\langle\alpha^{(l)}|) Z \right) - y_j^{(l)} \right| \\ \textit{subject to} \quad & Z \geq 0 \end{aligned} \tag{3.114}$$

and let $M_j^{*\sharp}$ *be the complex conjugate of the eigenvector of* $Z_j^\sharp$ *corresponding to its largest eigenvalue rescaled to have length* $\left\| M_j^{*\sharp} \right\|_{\ell_2} = \sqrt{\|Z_j^\sharp\|_\infty}$*. Then, we estimate* M *by*

$$M^\sharp = \begin{pmatrix} {M^\sharp}_1{}^T \\ \vdots \\ {M^\sharp}_n{}^T \end{pmatrix}. \tag{3.115}$$

Note that Eq. (3.115) simply amounts to stacking the separately recovered row vectors $M_j^\sharp$. Now, a simple extension of Theorem 3.5 yields a similar performance guarantee for Protocol 3.11: Due to the similarity of the intensity measurements (3.112) for a single row M_j of the transfer matrix and the measurements assumed in Theorem 3.5, the latter guarantees recovery of said row with high probability by means of PhaseLift (3.12). In order to succinctly state the final result, we introduce some additional notation. Define the total noise at detector j (measured in ℓ_1-norm) to be

$$\epsilon_j^{\text{tot}} = \sum_{l=1}^{m} |\epsilon_j^{(l)}| \tag{3.116}$$

and the overall noise strength:

$$\epsilon^{\text{tot}} = \sqrt{\sum_{j=1}^{n} {\epsilon_j^{\text{tot}}}^2}. \tag{3.117}$$

This formulation allows for treating the different output modes and their detector noise levels individually. In particular, we do not require a universal type of noise for all detectors, but allow for taking into account detector dependent noise of different strength, i.e. varying noise levels.

Corollary 3.12 (Performance guarantee for Protocol 3.11)**.** *The reconstruction* $M^\sharp$ *of any transfer matrix* M *by means of Protocol 3.11 satisfies*

$$\min_{\mu : |\mu_j| = 1} \left\| M^\sharp - \operatorname{diag}(\mu_1, \ldots, \mu_n) M \right\|_2 \leq C \frac{n \epsilon^{\text{tot}}}{m \nu}. \tag{3.118}$$

with probability at least $1-\mathcal{O}\left(\mathrm{e}^{-\gamma m}\right)$*. Here,* C *and* γ *are positive constant of sufficient size and*

$$\nu = \min_{1\leq j\leq n} \|M_j\|_{\ell_2}. \tag{3.119}$$

Recall that $\mathrm{diag}(\mu_1,\ldots,\mu_n)$ are the row-phases of M, which are unrecoverable from the intensity measurements (3.4). We include the additional correction (3.119) to deal with possible loss. Unitary transfer matrices satisfy $\nu = 1$.

Proof. For any fixed row vector M_j,

$$\min_{0\leq\phi\leq 2\pi} \left\| M_j^\sharp - \mathrm{e}^{i\phi} M_j \right\|_{\ell_2} \leq C' n \min\left\{ \|M_j\|_{\ell_2}, \frac{\epsilon_j^{\mathrm{tot}}}{m\|M_j\|_{\ell_2}} \right\}. \tag{3.120}$$

follows directly from Theorem 3.5. Note that the additional n factor compared to Eq. (3.61) is due to the normalization of the input vectors. The input vectors need to be scaled by $\sqrt{n}$ in order to be able to apply Theorem 3.5.

Before we can move on to determine the remaining row vectors M_i $(i\neq j)$ of M, it is important to point out that the recovery guarantees of Theorem 3.5 are *universal*: one instance of randomly chosen measurement vectors suffices to recover *any* vector $x\in\mathbb{C}^n$. This allows for applying this reconstruction guarantee to all n row vectors M_j simultaneously. The total noise bound (3.118) now follows from the entry-wise definition of the Frobenius norm:

$$\begin{aligned}
\min_{\mu} \left\| M^\sharp - D(\mu)M \right\|_2^2 &= \min_{0\leq\phi_1,\ldots,\phi_n\leq 2\pi} \sum_{j=1}^n \left\| M_j^\sharp - \mathrm{e}^{i\phi_j} M_j \right\|_{\ell_2}^2 && (3.121)\\
&= \sum_{j=1}^n \min_{0\leq\phi_j\leq 2\pi} \left\| M_j^\sharp - \mathrm{e}^{i\phi_j} M_j \right\|_{\ell_2}^2 && (3.122)\\
&\leq C^2 n^2 \sum_{j=1}^n \min\left\{ \|M_j\|_{\ell_2}^2, \frac{\eta_{(j)}^2}{m^2\|M_j\|_{\ell_2}^2} \right\} && (3.123)\\
&\leq (Cn)^2 \sum_{j=1}^n \frac{\eta_j^2}{m^2\|M_j\|_{\ell_2}^2} && (3.124)\\
&\leq \frac{(Cn)^2}{m^2\nu} \sum_{j=1}^n \eta_j^2 && (3.125)\\
&= \left(Cn\frac{\eta^{\mathrm{tot}}}{m\nu} \right)^2, && (3.126)
\end{aligned}$$

Here, we have used Eq. (3.61) for each summand in Eq. (3.123). □

The performance guarantee above has an interesting consequence for experimental design: The right hand side of Eq. (3.118) is mainly determined by the signal-to-noise ratio $\frac{\nu}{\epsilon^{\mathrm{tot}}/m}$. Remarkably, the noise term does not become smaller for increasing m. To be more precise, let us assume that each detection error is independent and normally distributed with standard deviation σ, i.e. $\epsilon_j^{(l)} \sim \mathcal{N}(0, \sigma^2)$. Then,

$$\mathbb{E}\epsilon_j^{\mathrm{tot}} = \frac{1}{m}\sum_l \mathbb{E}\epsilon_j^{(l)} = \mathbb{E}\epsilon_j^{(1)} = \sqrt{\frac{2}{\pi}}\sigma, \tag{3.127}$$

and hence, the expected error $\mathbb{E}\epsilon^{\mathrm{tot}}$ is independent of m. Of course, the standard deviation of $\epsilon_j^{\mathrm{tot}}$ scales as $\frac{1}{\sqrt{m}}$. Therefore, an increase of m past the threshold Cn in Corollary 3.12 mainly influences the (exponentially small) failure probability. In other words, said corollary implies that once the sampling threshold is reached, there is not much use to further increase the number of measurements as the error bound (3.118) is primarily determined by the uncertainties of a single measurements. Hence, any additional experimental time should be invested to reduce the uncertainty of the single measurements, e.g. by increasing the number of single photon events used to estimate $I_j(\alpha)$.

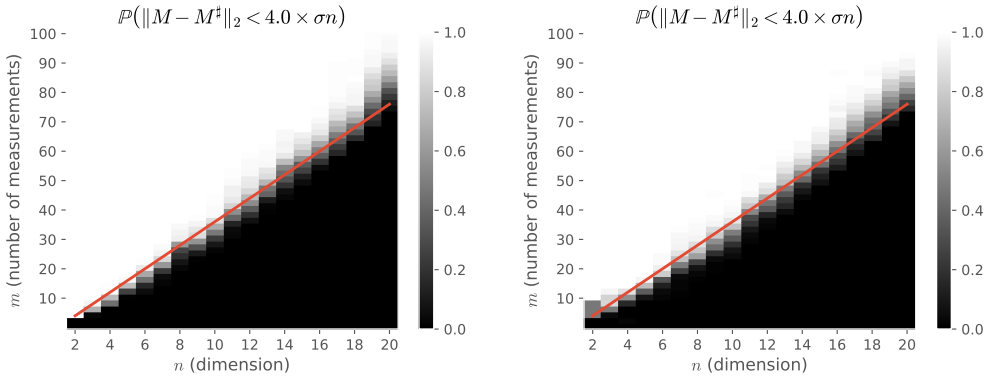
By studying the assumptions and the proof of the underlying Theorem 3.5, we also note that this behavior is to be expected. Since this theorem does not assume any statistical properties of the noise, but only assumes that the noise $\epsilon^{(l)}$ is bounded, we cannot expect a statistical improvement by increasing the number of measurements. For example, if we assume a constant, purely systematic error $\epsilon^{(l)} = c$ for all measurements, then no improvement can be expected even for $m \to \infty$.

However, it should be kept in mind that Corollary 3.12 only provides necessary conditions for recovery, which are not optimal in the large m regime. Therefore, we perform numerical simulations in the next chapter to further investigate how experimental time should be spent. In other words, we study the question how the reconstruction performs as a function of m when the total experimental time budget is fixed. Note that this behavior has already been explored in the context of quantum state tomography via compressed sensing in [Fla+12].

3.4. Application

3.4.1. Numerical results

We demonstrate the practical applicability of the PhaseLift characterisation protocol using simulated experiments. The simulation depicted in Fig. 3.2 aims to visualize the performance guarantees from Corollary 3.12: For each given dimension n, we choose 100 target unitaries. Each of these is reconstructed by means of Protocol 3.11 with a varying number of measurements m. The input vectors are sampled from the



(a) Uniform sampling

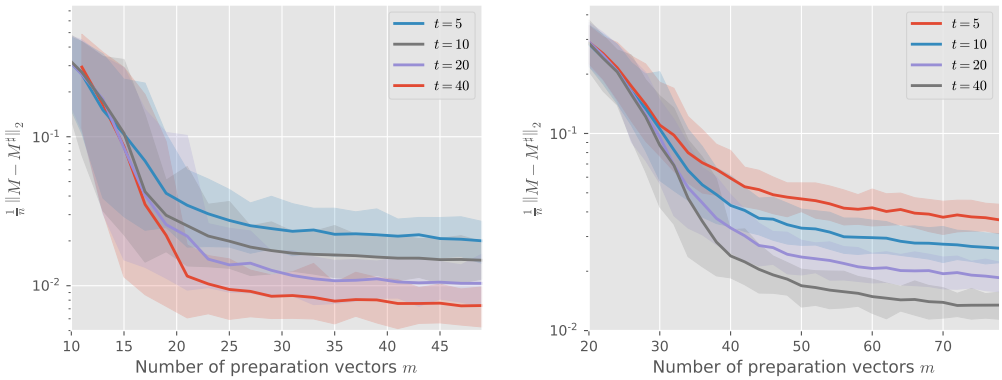
(b) RECR sampling

Figure 3.2.: Simulated recovery-probability using the two different sampling schemes under noisy measurements with $\sigma = 0.05$. For each given dimension, the transfer matrices to be recovered consist of 97 Haar random unitaries as well as the identity, the swap-matrix, and the discrete Fourier transform. The red line indicates the conjectured phase transition at $4n - 4$.

uniform ensemble in Fig. 3.2a and from the normalized RECR ensemble in Fig. 3.2b. For the measurement noise ϵ_j from Eq. (3.112), we assume independent, centered Gaussian noise with standard deviation $\sigma = 0.05$. The density plots show the fraction of successfully recovered unitaries. Here, the criterion for success is whether the distance of the reconstruction $M^\sharp$ measured in Frobenius norm is smaller than the threshold $4\sigma n$ in accordance with the error bound (3.118).

Figures 3.2a and 3.2b show a pronounced phase transition around $m = 4n$. This demonstrates the high sample efficiency of the PhaseLift reconstruction. Not only does the number of measurements scale linearly in the system size – as rigorously proven in Corollary 3.12 – but the scaling coefficient is small as well.

In the simulations depicted in Fig. 3.2, we assumed a constant noise level for each m. Therefore, the lab-time required for taking the data or, put differently, the number of single photon events required for estimating the intensities increases linear in m. We now investigate the question posed at the end of Section 3.3.3, namely how the reconstruction performs as a function of m when the total experimental time budget is fixed. Recall from Eq. (3.8) that the single photon counting statistics is given by a multinomial distribution with number of trials N given by the total photon number and the probabilities p_j given by the expectation values in Eq. (3.7). Denote by $N^{(l)}$ the number of photons used to estimate the output intensities for a single photon input state $|\psi(\alpha^{(l)})\rangle$ with $l = 1, \ldots, m$. In Fig. 3.3, we depict the reconstruction error



(a) $n = 5$ (b) $n = 10$

Figure 3.3.: Simulated reconstruction error using RECR measurements for a fixed time budget as a function of the number of distinct input vectors for two different circuit sizes and different time budgets. The total photon number budget for each reconstruction is $N = \Gamma \times t$, where $\Gamma = 4600\text{s}^{-1}$ is a typical counting rate of the experiment and t the time available. Then, the output for each input vector α is a multinomial distribution with the number of trials given by $\frac{N}{m}$. We sample 100 sets of measurement data for each value of m and run the PhaseLift reconstruction on each of them. The solid line indicates the mean error and the colored areas the 0.025 and 0.975 quantiles.

with the total number of photons used for reconstruction $N = \sum_l N^{(l)}$ kept fixed. To be more precise, we choose the total number of photons N as a multiple of the counting rate from the experiment $\Gamma = 4600\, s^{-1}$ introduced in Section 3.1, i.e.

$$N = t \times \Gamma, \tag{3.128}$$

where t is the time spent only on taking the single-photon data. Therefore, we have $N^{(l)} = \frac{N}{m}$, where m is the number of preparation vectors. The reconstructions in Fig. 3.3 are then performed by randomly sampling 100 outcomes from the output counting statistics of each input state. For larger m, the $N^{(l)}$ becomes smaller, and therefore, the statistical error in each estimated intensity grows.

First, we see from Fig. 3.3 that – as expected – taking more data by increasing t improves the overall reconstruction quality. Also note that the reconstruction error is approximately independent of m above a certain threshold. This clearly shows that the recovery guarantee in Corollary 3.12 is not tight for larger values of m, as the right hand side of Eq. (3.118) grows with the individual statistical error of each measurement.

From Fig. 3.3, one could conclude that there is no advantage of taking a small value of m in the experiments. This conclusion rests on the assumption that the total number of photons is the figure of merit that best describes an experimentalist's budget. However, in the concrete experimental architecture introduced in Section 3.1, this is not the case: In reality, the number of distinct settings for the reconfigurable chip is more critical for the time required to perform a given experiment. This is due to the fact that switching the reconfigurable phase shifters and couplers takes more time that the actual data taking process. Therefore, if we take into account this additional cost for reconfiguring the preparation stage on the chip, reconstructions with a smaller number of more precise measurements perform better when the "time budget" is kept fixed.

3.4.2. Experimental results

To demonstrate its practical utility, we use the PhaseLift protocol to perform experimental reconstruction on the reconfigurable integrated photonic circuit introduced in Section 3.1 In Fig. 3.4, we show the reconstruction error of the PhaseLift approach. Since our aim is to benchmark the performance of the characterization technique, and not the performance of the chip itself, we compare to other reconstructions obtained through established but more costly techniques: For the smaller transfer matrices of dimension two and three, we perform a complete HOM-dip-reconstruction based on two photon interference as described in Appendix A.2.1. However, since this is infeasibly costly for the five-dimensional transfer matrices, we only compare these to single-photon reconstructions of the absolute values of the transfer matrix components. For more details, see Appendix A.2.

The number of input vectors used in each PhaseLift reconstruction is $m = 5n$. This slight overhead compared to the conjectured and numerically observed phase transition in Fig. 3.2 is used to counteract systematic errors in the preparation of the input vectors. To be more precise, we conjecture that the main source of error in this experiment is due comparatively large errors in preparation of certain input states. By increasing m slightly, these effects average out and provide in total a smaller reconstruction error.

We see that the PhaseLift reconstructions and the references agree well for most settings in Fig. 3.4 Even without exploiting the possible advantages of the RECR ensemble due to a better calibration, it generally performs as well as the uniform ensemble. Both display a similar behavior: for a fixed number of modes, the deviations are generally larger for the random unitaries compared to the more structured identity and Fourier transfer matrices. The errors for the two-dimensional transfer matrices are slightly smaller than for the corresponding three-dimensional transfer matrices as expected from Eq. (3.118). Furthermore, the currently used sequential arrangement of the Mach-Zehnder interferometers in the preparation stage of the

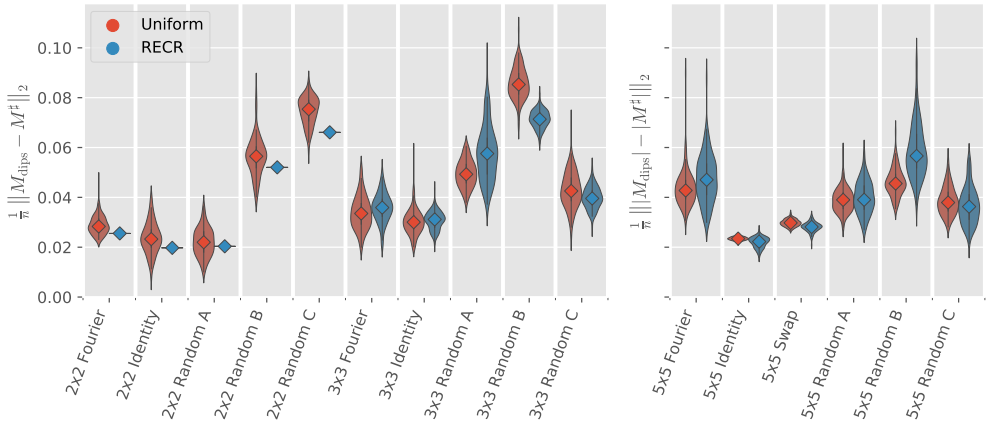


Figure 3.4.: Comparing reconstructions from experimental data for different target transfer matrices and sampling schemes. For each matrix and sampling scheme, we subsample $m = 5n$ preparation vectors and the corresponding measured intensities from the experimental data 100 times. To estimate the quality of the reconstruction, we plot the discrepancy between the PhaseLift reconstruction and an alternative method. The diamonds indicate the median and the colored area sketches the distribution of this reconstruction discrepancy. In the left picture, the reference is obtained through a HOM-dip reconstruction as discussed in the appendix. However, since this technique is too costly for larger dimensions, the five dimensional reconstructions on the right are only compared in magnitude to a reference from single photon data, which is observation to all phase information. Since for $n = 2$ there are only six distinct RECR vectors up to a global phase, only the median is shown in these cases. For more details on the data analysis see the supplemental material.

experiment leads to higher deviations of the actual prepared compared to the intended measurement vectors with an increase in n. Of course, this leads to larger reconstruction errors as well – possible solutions to this problem are discussed in the conclusions. For the reference reconstruction, we also expect larger deviations with an increase in the size of the transfer matrix, since errors from reconstructing one row accumulate in the error for other rows as well. Note that the errors for the five-dimensional transfer matrices are relatively small since they only take into account the absolute values of the components and neglect all phases.

In Fig. 3.5, we directly compare the performances of the reconstruction protocols, namely of the PhaseLift reconstruction and the HOM-dip reconstruction. In contrast

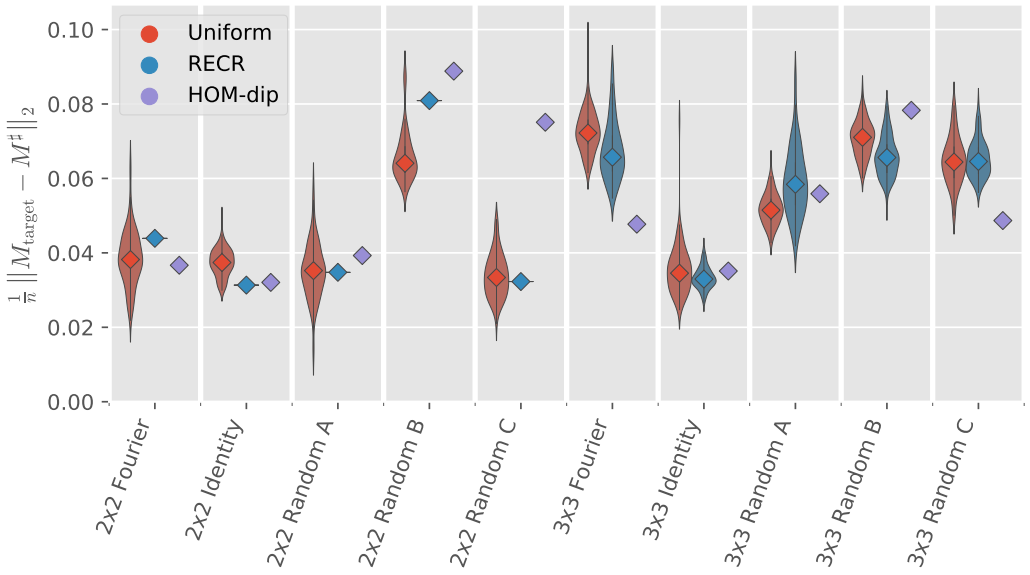


Figure 3.5.: Same as Fig. 3.4, but the reconstructions are compared to the theoretical target unitaries. "HOM-dip" refers to the reconstructions used as references in Fig. 3.4. We do not show the results for the 5 dimensional unitaries since the corresponding HOM-dip reconstructions were too costly to take.

to Fig. 3.4, we use the theoretical target unitary as a reference. Generally, the errors of the PhaseLift reconstructions and the HOM-dip reconstructions are of the same order of magnitude. This is despite the fact that the HOM-dip reconstruction is not just insensitive to the row phases, but also to the column phases. Therefore, the reported errors for the HOM-dip reconstruction are minimized over both row- and column phases instead of just the row phases for the PhaseLift reconstruction. The additional free parameters in the minimization may lead to overfitting, and hence, to an underestimation of the actual error of the HOM-dip reconstruction.

Finally, we study the influence of varying the number of measurements and the statistical error on each measurement in Fig. 3.6. In the left picture, we randomly select m measurements from the existing data for 100 times and plot the mean as well as the spread of the error. Except for a larger error for very few measurements, the RECR ensemble performs equally well as the Gaussian ensemble. We see that the error saturates at a non-zero value, which might be due to systematic errors. Also, note that the reconstruction error already saturates around $m = 4n = 20$

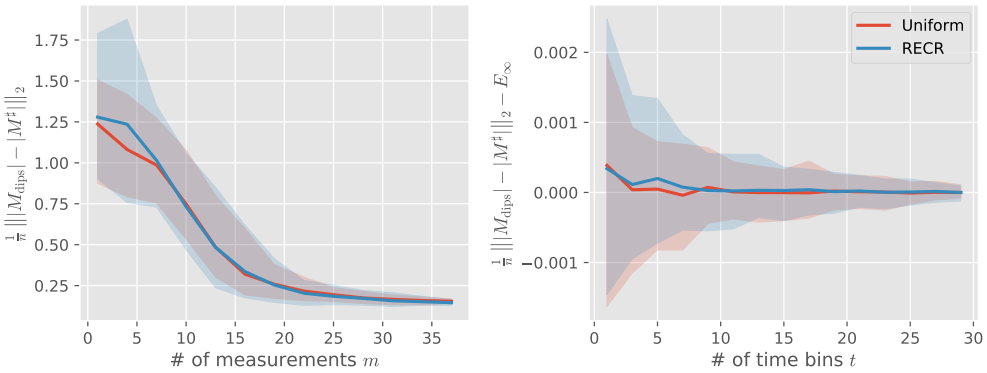


Figure 3.6.: Reconstruction errors for a random 5×5 transfer matrix from experimental data. For each picture, we plot the mean (solid) as well as the 0.025 and 0.975 quartiles over 100 samples. In the left picture, each sample consists of a recovery from m preparation vectors and the corresponding photon counts measured over 30 s In the right picture we fix a randomly selected set of $m = 20$ preparation vectors and run the recovery with the photon counts from t randomly selected time bins, each of which is one second long. The constant shift E_∞ in the right picture is equal to the mean error for $t = 29$ and it serves the purpose to equalize any differences between the Gaussian and the RECR reconstructions due to the choice of preparation vectors.

measurements.

To further investigate the source of the reconstruction error, we vary the statistical error of the measurements $\epsilon^{(l)}$ in the right picture by changing the number of photons used to estimate $I_j(\alpha^{(l)})$ in terms of Eq. (3.7). Since the experimental data consists of photon counts per one second time bin, we can vary the number of total photons by randomly selecting t time bins per reconstruction. In the right hand picture of Fig. 3.6, we fixed a random subset of 20 measurements and observe the reconstruction error as a function of t. Furthermore, we normalized the reconstruction error with a constant offset E_∞ for each measurement scheme to cancel out reconstruction errors due to the different choice of measurement vectors. As we can see in the left image, the spread of the reconstruction error for $m = 20$ is still on the order of ±0.1, and therefore, dominating the smaller error due to statistical fluctuations. We chose E_∞ such that the mean error for the largest value of t is 0. First, we see that the reconstruction error due to statistical uncertainty is small compared to the total error even for very small values of t. This is due to the large counting rates of modern single photon experiments, which is $\Gamma \approx 4600\,\mathrm{s}^{-1}$ in this case. Therefore, the main

source of error in this experiment is due to systematic errors such as an relatively inaccurate preparation of the vectors $\alpha^{(l)}$. We already mentioned that this can be rectified by better calibrating the preparation stage, which is much easier to perform for the RECR ensemble than for the uniform ensemble.

3.5. Conclusion & outlook

In this chapter, we introduce a characterization technique for linear optical networks based on this problem's connection to phase retrieval. First and foremost, we show that recovering each row of the transfer matrix constitutes a separate phase retrieval problem, which can be solved using convex programming. Provided that the input vectors are chosen and random from a suitable distribution, the number of distinct inputs required scales linear in the number of modes. To prove this statement, we exploit the rank-one constraint, which arises from lifting the quadratic phaseless measurements to linear measurements on a larger spaces.

Motivated by the specific application in linear optics, we propose the RECR measurement ensemble. The main technical contribution of this work is the proof that despite being highly structured, the RECR ensemble performs just as well for phase retrieval as the well-established uniform/Gaussian measurement ensembles. More generally, we introduce the notion of (super)-attentive measurement ensembles and show their efficacy for efficiently reconstructing rank-one matrices from few measurements.

Finally, we report on ongoing experimental work to implement the proposed reconstruction protocol using a small-scale universal optics chip. Our preliminary results show that the PhaseLift characterization technique is a valuable tool for experimentalists due to its sample efficiency and robustness to noise.

Future work on phase retrieval may benefit from the general reconstruction guarantees proven in this work as the conditions for (super)-attentiveness are quite general. The same statement applies to the majorization trick, which we use to generalize the proofs from attentive to super-attentive ensembles. The question remains whether any other practically relevant measurement ensembles fall within this framework.

Furthermore, the current work is restricted to the reconstruction of rank-one matrices. We are interested in the question whether the RECR ensemble can be used for low-rank matrix recovery in general as the Gaussian ensemble [KRT17]. Matrix multiplications with structured measurements are often computationally faster, and therefore, structured ensembles are also interesting outside of particular applications.

For the specific application of characterizing linear optical networks, future work should investigate the applicability of our approach to larger systems, which may be

out of reach for other techniques. One of the main problems that need to be solved for this is a more accurate preparation stage: The preparation procedure used so far has not explicitly made use of the advantages of the RECR ensemble, that is the finite number of phases and magnitudes, which need to be prepared in each mode. Another drawback of the current architecture is that possible errors in the preparation state for the coherent state inputs $|\alpha\rangle$ add up linearly due to the serial wiring. To circumvent these two problems, we propose to investigate a tree-like arrangement of the directional couplers. This way, we could attain more independence of the phase shifters and couplers.

4. Low-rank tensor recovery

The theoretical and numerical study of many-body quantum systems is severely hindered by the *curse of dimensionality*, which in this context refers to the exponential growth of the dimension of the Hilbert space of quantum states w.r.t. the number of constituents. However, many realistic systems do not exhibit this pathological complexity and can be described efficiently in terms of *tensor networks* [BC17; Orú14]. Here, "efficient description" refers to the fact that the number of parameters required to describe such a state scales polynomially in the number of subsystem. An especially simple but important special case of tensor networks is known under names such as *finitely correlated states* [FNW92] or *matrix-product states* (MPS) [Per+07; VMC08]. Many states used in quantum information processing have an efficient description in terms of an MPS.

The idea of low-complexity tensor representations have also received attention in the context of machine learning. One important application is model compression: State of the art deep neural networks have millions of parameters and high-dimensional hidden layers, which complicates deploying them on devices with limit capabilities such as smart phones or IoT devices. By "compressing" the learned weights using efficient tensor formats, we can reduce the complexity of the model with only slight degradation in performance [Nov+15; Tai+15; YH16].

The question we are attempting to answer in this chapter is whether such MPS can be efficiently reconstructed from few linear measurements[1]. More precisely, we provide analytical and numerical evidence that the number of measurements required to recover an MPS is related to its intrinsic complexity – and hence, scales polynomially in the number of constituents – although the Hilbert space of tensors grows exponentially in the number of constituents. This work is a natural extension of *low-rank matrix recovery*, which formed the foundation of the results in Chapter 3, to higher-order tensors.

This chapter is structured as follows: In Section 4.1, we introduce the MPS tensor format and in Section 4.2, we present the software library MPNUM dealing with tensors in MPS representation. MPNUM was developed as part of this work to facilitate numerical computations in a user friendly and reusable manner and is the foundation

[1] Note that since we assume a linear measurement model, the results of this section do not apply to the problem of estimating pure quantum states. We use the term "matrix product state" to not only refer to quantum states, but to general tensors in the MPS format.

of all the numerical experiments in this chapter. Finally, Section 4.3 reports on work in progress, which is concerned with provably recovering MPS from few linear measurements using an *Alternating Least Squares* (ALS) algorithm.

Relevant publications

- Ž. Stojanac, D. Suess, M. Kliesch: *On the distribution of a product of N Gaussian random variables*, Proceedings Volume 10394, Wavelets and Sparsity XVII; 1039419 (2017)
- Ž. Stojanac, D. Suess, M. Kliesch, *On products of Gaussian random variables*, arXiv:1711.10516
- D. Suess, M. Holzaepfel, *mpnum: A matrix product representation library for Python*, Journal of Open Source Software, 2(20), 465 (2017)

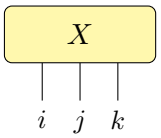
4.1. Matrix Product States

Tensors are a generalization of vectors and matrices. Although there is a coordinate-free definition for tensors in terms of multi-linear functionals [Bro12], we are going to identify a tensor with its coordinate representation w.r.t. a fixed basis here. A complex tensor of order N is an element $X \in \mathbb{C}^{d_1 \times \cdots \times d_N}$, where the d_i are called *local dimensions*. Hence, a vector is a tensor of order 1 and a matrix is a tensor of order 2. For the sake of simplicity, we assume that $d_1 = \cdots = d_N$ throughout this chapter.

4.1.1. Graphical notation

Since formulas with higher-order tensors may become incomprehensible due to the large amount of indices, we introduce a widely used graphical notation [Orú14; BC17] here. A tensor $X \in \mathbb{C}^{d^N}$ is represented by a geometric shape with legs attached, where each leg corresponds to one index. For example, consider the case $N = 3$, then the components of the tensor X are given by

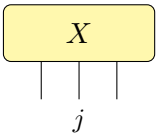
$$X_{i,j,k} = \quad \text{(4.1)}$$



A variable written next to a leg fixes the corresponding index to the given value, while an unlabeled tensor leg represents an unmatched index. In the following, we make use of the following notation for an unmatched index inspired by Python's

and Matlab's syntax: As an example, consider X as above, then we define the slice $X_{:,j,:} \in \mathbb{C}^{d^2}$ for each j by

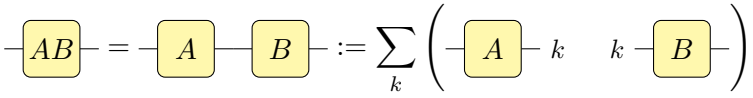
$$X_{:,j,:} = \quad . \tag{4.2}$$



Written out explicitly, we have the slightly cumbersome equality $(X_{:,j,:})_{i,k} = X_{i,j,k}$. The advantage of this graphical notation becomes clear once we express tensors composed of other tensors by operations such as contraction or tensor products. The most common operations and their graphical notation are summarized below:

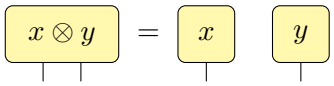
- **Contractions** are indicated by joining two legs, e.g. for matrices $A, B \in \mathbb{C}^{d\times d}$, their product AB is written as

$$\quad = \quad := \sum_k \left(\quad k \quad k \quad \right) \tag{4.3}$$



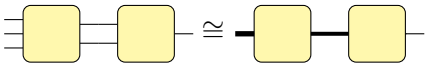
- **Tensor products** correspond to drawing two tensors side by side, e.g. for $x, y \in \mathbb{C}^d$ their tensor product $x \otimes y$ is written as

$$\quad = \quad \quad \tag{4.4}$$



- **Grouping** indices – that is the canonical identification $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \cong \mathbb{C}^{d_1 d_2}$ – is indicated by merging two or more legs together.

$$\quad \cong \quad \tag{4.5}$$



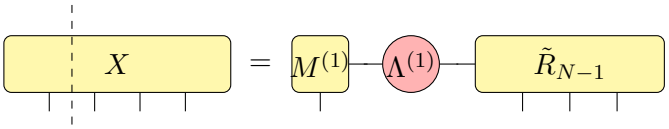
This operation is often used in numerical implementations of tensor network algorithms as it reduces most tensor operations to standard matrix operations. For example, the twofold contraction on the left hand side of Eq. (4.5) is converted to a matrix multiplication on the right hand side. In the following, we often perform grouping on neighboring tensor legs implicitly.

4.1.2. MPS tensor representation

The tensor representation we are interested in has been established independently in quantum physics under the names *finitely correlated states* [FNW92] and *matrix-product states* [KSZ91; KSZ92]. However, similar structures have been known for much longer in the form of *hidden Markov models* [CMR06]. Subsequently, it has

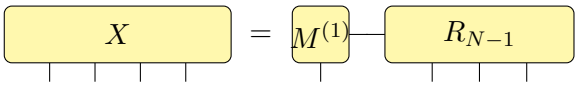
been rediscovered in other contexts: It is known under the guise of the *density matrix renormalization group* [Whi92; Sch11] in condensed matter physics. Furthermore, the MPS representation is also known as a special case of the *Hierarchical Tucker* tensor format [Hac12; Gra10] and under name *tensor-train* [Ose11] in the applied math community.

To introduce the MPS representation, consider a tensor $X \in \mathbb{C}^{d^N}$. By splitting off the first index, grouping the remaining $N-1$ indices, and performing a SVD on the resulting matrix, we can factor X into three parts



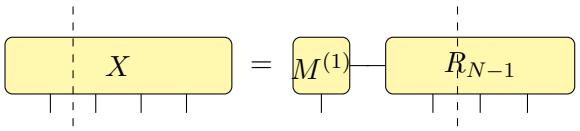
(4.6)

where $M^{(1)}$ denotes the left-singular vectors, $\Lambda^{(1)}$ the diagonal matrix composed of the singular values $\lambda^{(i)}$, and $\tilde{R}_{N-1}$ the left-singular vectors after the index-grouping has been reversed. For convenience, we contract the singular values with the remainder $\tilde{R}_{N-1}$ and obtain

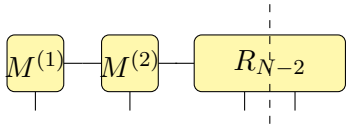


(4.7)

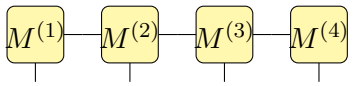
This procedure can now be iterated by performing the same steps on R_{N-1}:



(4.8)

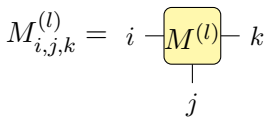


(4.9)



(4.10)

This yields a representation of X in terms of the *local tensors* $M^{(l)}$. The above algorithm for computing the local tensors is referred to as TT-SVD in [Ose11]. In the following, we identify the index order as follows:



(4.11)

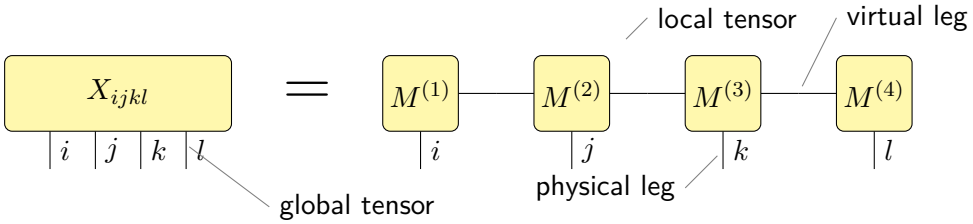


Figure 4.1.: A 4th order tensor in MPS representation with open boundary condition as described in Eq. (4.13).

The horizontal legs corresponding to the indices i and k are often referred to as *bonds* or *virtual legs*, and the vertical ones corresponding to j are referred to as *physical* legs. By fixing each physical leg to some value k_l in Eq. (4.10), we see that each component of X is given by product of N matrices

$$M^{(l)}_{k_l} := M^{(l)}_{:,k_l,:}, \tag{4.12}$$

hence the name "matrix-product representation". In this case, we have

$$X_{i_1,\dots,i_N} = M^{(1)}_{i_1} \cdots M^{(N)}_{i_N}, \tag{4.13}$$

where we used the shorthand notation from Eq. (4.12)

So far this construction is exact and completely general, i.e. every tensor can be represented in MPS form. However, this representation still requires exponentially many parameters in general: Generically, the SVD in the i-th step yields

$$r_i = \min d^i, d^{N-i} \tag{4.14}$$

non-zero singular values $\lambda^{(i)}$, and therefore, the local tensors $M^{(i)}$ have shape (r_{i-1}, d, r_i). However, suppose that all but r of the singular values in each cut vanish, then we can truncate the local tensor $M^{(i)}$ to shape (r, d, r) while still preserving the exact equality (4.10). In this case, the number of parameters scales as $\mathcal{O}(r^2 N d)$ and X can be represented efficiently. Although this assumption may seem overly strict on a first glance, many tensors in our applications of interest are well approximated by such an MPS with small r. The following definition sums up the discussion in this section so far.

Definition 4.1. *A* matrix-product state representation *of a tensor $X \in \mathbb{C}^{d^N}$ is defined in terms of a tuple of local tensors $(M^{(l)})_{l\in[N]}$ with $M^{(l)} \in \mathbb{C}^{r_{l-1}\times d\times r_l}$ where $r_0 = r_N = 1$. Then, the components of X are given by*

$$X_{i_1,\dots,i_N} = M^{(1)}_{:,i_1,:} \cdots M^{(N)}_{:,i_N,:}. \tag{4.15}$$

We say that the representation in Eq. (4.15) *has* open boundary conditions.

By allowing larger boundary ranks r_0 and r_N, and tracing over the resulting matrix in Eq. (4.15), we obtain MPS with periodic boundary conditions. These are more suited if there are strong correlations between the first and the last site as well, e.g. if the geometry of subsystems is not linear but circular. As already noted above, the crucial parameter balancing the efficiency of the variational class of MPS on one hand and its expressive power on the other hand is the size of the virtual legs r_l. This warrants the following definition.

Definition 4.2. *Let $(M^{(l)})_{l\in[N]}$ denote an MPS representation with $M^{(l)} \in \mathbb{C}^{r_{l-1}\times d\times r_l}$. We call $\max_l r_l$ the (MPS-)*rank *or* bond dimension *of the MPS. Furthermore, for any tensor X, we define its* MPS-rank *to be the smallest rank of any MPS representation of X.*

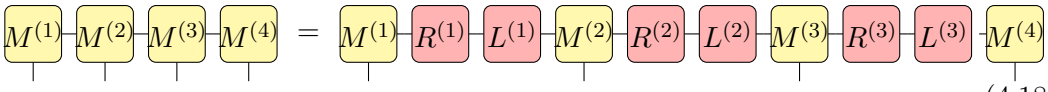
Note that by [Ose11, Thm. 2.2], the definition of MPS-rank in Definition 4.2 agrees with the usual definition using matrications of X, and hence, is well defined. In contrast to the matrix case, there are multiple inequivalent definitions rank for tensors [KB09], which arise from different tensor representations. Throughout this work, we focus on the MPS-representation and the corresponding notion of rank due to its advantages outlined in Section 4.1.3 Therefore, we simply call the MPS-rank "rank".

With the factorization of the tensor X in Definition 4.1, we have introduced new virtual degrees of freedom such that the partial trace over them equals X. Clearly, the following gauge transformation leaves Eq. (4.13) invariant

$$\tilde{M}_i^{(1)} = M_i^{(1)} R^{(1)}, \tilde{M}_i^{(N)} = R^{(N-1)} M_i^{(N)}, \tag{4.16}$$

$$\tilde{M}_i^{(l)} = L^{(l-1)} M_i^{(l)} R^{(l)} \quad (1 < l < N) \tag{4.17}$$

provided the (generally non-square) matrices $L^{(l)}$ and $R^{(l)}$ satisfy $L^{(l)}R^{(l)} = \mathbb{1}$ since then

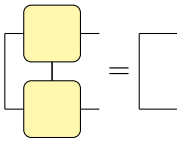


(4.18)

As shown in [Per+07, Thm. 2], these local transformations on the virtual degrees of freedom are the only possible gauge transformations of the MPS representation. A number of canonical forms exist that partially fix the gauge [Per+07; Sch11; BC17]. The following definition summarizes the corresponding gauge conditions.

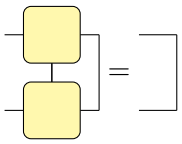
Definition 4.3. *We say that a local tensor $M^{(l)}$ is* left-normalized *if it satisfies*

$$\sum_k M_k^{(l)\dagger} M_k^{(l)} = \mathbb{1} \qquad \text{=} \tag{4.19}$$



A MPS with all but the rightmost local tensors in left-normalized form is called left-canonical. *Similarly, a* right-normalized *local tensor fulfills*

$$\sum_k M_k^{(l)} M_k^{(l)\dagger} = \mathbb{1} \qquad (4.20)$$

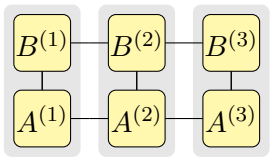


and a right-canonical *MPS has all local tensors in right-normalized form except for the first.*

The exceptions for the local tensors at the beginning and end of the chain are necessary to accommodate tensors with Frobenius norm different from one. In practice, we often deal with a mixed left-right-canonical form, e.g. in variational algorithms updating one local tensor at a time such as DMRG or the alternating least squares algorithm presented in Section 4.3. For these purposes, efficient algorithms exist that transform an MPS to a canonical form [Sch11; Orú14]. If we set aside transformations that change the bond dimensions, the remaining gauge group is unitary, and hence, computations on canonical MPS are far more stable numerically.

We have already emphasized the ability of the MPS representation to represent certain tensors efficiently, i.e. with a number of parameters scaling polynomially in the order of the tensor. One crucial advantage of this tensor format is that it also facilitates efficient arithmetical operations for tensors [Sch11; Orú14]. In other words, operations such as sums and contractions of MPS are also MPS and the corresponding local tensors can be computed efficiently from the local tensors of the inputs. For example, consider the scalar product of two tensors $A, B \in \mathbb{C}^{d^N}$:

$$\langle A, B \rangle = \qquad (4.21)$$



By contracting the physical legs of the local tensors first as indicated by the gray boxes[2], we obtain an MPS representation of a tensor of 0th order, i.e. a scalar, which can be contracted afterwards to compute its value.

As most arithmetic operations increase the virtual dimension substantially, we need methods to approximate the result by another MPS with lower virtual dimension while keeping the approximation error small. Although computing the *best* rank r approximation of an MPS is computationally infeasible in general [HL13], there are

[2]Note that this strategy is not optimal in many cases [Sch11] but suffices as an example here.

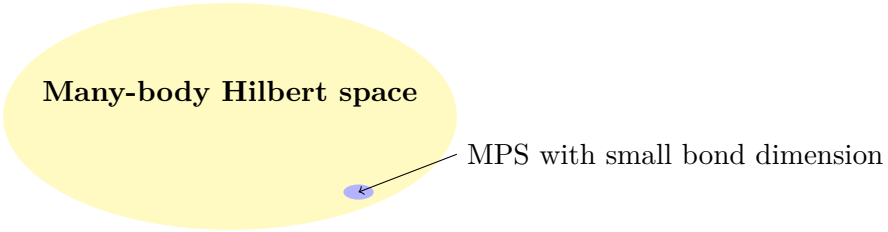


Figure 4.2.: The manifold of quantum states with an efficient MPS description occupies only a tiny corner in the full Hilbert space.

efficient algorithms that produce good results in practice: One of them is *SVD compression*, which successively performs an SVD followed by a truncation of negligible singular values on the local tensors similar to the higher-order SVD (4.10) [Sch11]. If we truncate all but the r largest singular values in each step, we obtain a quasi-optimal rank r approximation of $X \in \mathbb{C}^{d^N}$, i.e.

$$\|X - X_{\mathrm{SVD}}\| \leq \sqrt{N-1}\|X - X_{\mathrm{best}}\|. \tag{4.22}$$

Here, $\|\cdot\|$ denotes the Frobenius norm of tensors, X_{SVD} the rank r SVD compression, and X_{best} the best rank r approximation of X.

4.1.3. Applications of the MPS format

The main application of the MPS representation in quantum information and condensed matter physics is the efficient description of certain many-body states. As the corresponding Hilbert space grows exponentially fast in the number of constituents N, the full description of any state in terms of its coefficients w.r.t. a fixed basis is only feasible for small N. Fortunately, not all quantum states of a many-body system are equally relevant in practice. Many systems of interest are well described by a Hamiltonian with local interactions, e.g. nearest neighbor interactions, which reflects in the structure of correlations in their low energy spectrum. More specifically, low energy eigenstates of gapped Hamiltonians with local interactions obey the *area-law* for entanglement entropy [Has06; VC06; ECP10]: For those states, the entanglement entropy of a subsystem asymptotically only depends on its boundary size and not on its volume. Although those stats occupy only a small corner of the enormous many-body Hilbert space as depicted in Fig. 4.2, they are exceptionally important in practice.

In one dimensional systems, any many-body state $|\psi\rangle$ with an area law[3] has an efficient representation. If we expand $|\psi\rangle$ in a product basis with coefficient tensor C,

[3] In one dimension, the boundary area of any connected region is constant and independent of the size of the region, and therefore, the area law implies constant entanglement entropy across bipartitions.

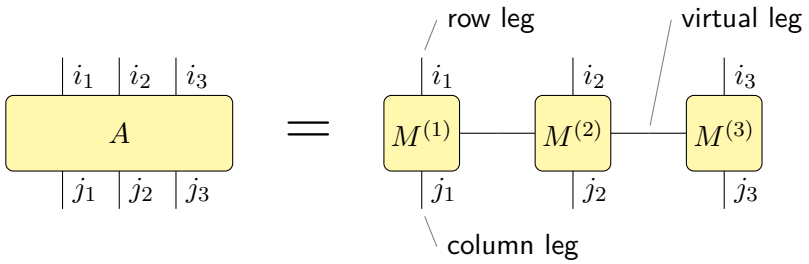


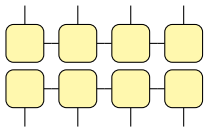
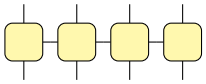
Figure 4.3.: An MPO with open boundary conditions.

$$|\psi\rangle = \sum_{i_1,\ldots,i_N} C_{i_1,\ldots,i_N} |i_1\rangle \otimes \cdots \otimes |i_N\rangle, \tag{4.23}$$

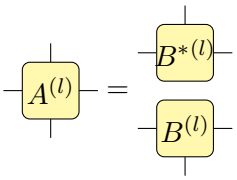
then C can be efficiently approximated by an MPS with bond dimension scaling polynomially in N and the inverse approximation error [Has06; VC06; ECP10; Ara+13; Ara+17]. States of the form (4.23) are appropriately referred to as matrix product states.

For higher dimensional regular latices, one can generalize the MPS representation. The resulting efficient tensor network representation is referred to as *projected entangled pair states* (PEPS). However, a statement analogous to the one above does not hold: For dimensions $D > 1$, not every state satisfying an area law can be efficiently represented as a PEPS [GE16] .

So far, we were only concerned with multi-body pure states. The MPS tensor format from Section 4.1.2 can also be adapted for the description of mixed sates. As an example consider an MPS[4] $|\psi\rangle$. The corresponding pure state projector can be written as

$|\psi\rangle\langle\psi| =$  $=:$  (4.24)

Here, the right hand side is a *matrix product operator* (MPO). Its local tensors $A^{(l)}$ are defined in terms of the local tensors $B^{(l)}$ of $|\psi\rangle$ as follows



(4.25)

[4]From now on we do not distinguish between the state and its corresponding coordinate representation w.r.t. a fixed product basis.

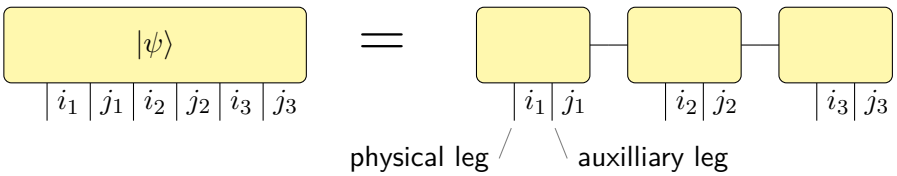


Figure 4.4.: An PMPS with open boundary condition.

with implicit grouping of the virtual indices of $B^{(l)}$. This is an example of a matrix-product density operator (MPDO) [VGC04; ZV04]. In general, any operator acting on the many-body Hilbert space can be expressed as an MPO with local tensors $A^{(l)} \in \mathbb{C}^{r\times d\times d\times r}$ as depicted in Fig. 4.3. This representation is efficient if the operator acts "locally", e.g. for a Hamiltonian with nearest neighbor interactions.

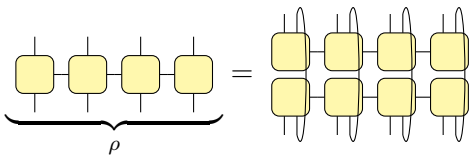
The explicit parametrization (4.25) of the local tensors makes it easy to see that the corresponding MPO is positive semidefinite (psd), and hence, represents a valid physical state. A crucial question for numerical computations with density operators expressed as MPO is if there is an efficient algorithm that decides whether the MPO is psd or not. Unfortunately, this problem is **NP**-hard, even if we restrict to dimensions $d = 2$ and translational invariant states [KGE14].

One way to cope with this problem is to use a manifest psd parameterization of the mixed state. Denote by $\mathcal{H}_1$ the Hilbert space of the system and by ρ the corresponding mixed state. By introducing a second *auxiliary* Hilbert space $\mathcal{H}_2$, we can write [NC10]

$$\rho = \mathrm{tr}_{\mathcal{H}_2} |\psi\rangle\langle\psi|, \tag{4.26}$$

where $|\psi\rangle \in \mathcal{H}_1 \otimes \mathcal{H}_2$. Note that we can take $\mathcal{H}_2$ to be of the same dimension as $\mathcal{H}_1$.

Similarly, a *local purification matrix product state* (PMPS) [De +13] is a matrix product state with two legs per site as depicted in Fig. 4.4 such that



(4.27)

This is a manifestly positive semidefinite parametrization of ρ in matrix product form. However, this advantage comes at a price: The PMPS representation of a state can be arbitrarily more costly than its MPDO representation [De +13]. More precisely, there a families of states $(\rho_N)_{N\in\mathbb{N}}$ with $\rho_N \in \mathbb{C}^{d^N}$ such that ρ_N has constant MPO-rank, but the PMPS-rank of ρ_N scales as $\mathcal{O}(N)$. This and the hardness result from [KGE14] shows that no efficient algorithm for computing an exact local purification from an

MPDO exists in general. However, the questions, how generic this hardness is and how hard approximative versions of this problem are, remain open.

Besides the original applications in physics, the MPS tensor format has recently also found applications in the field of machine learning: Deploying the latest neural networks on mobile devices such as smart phones or IoT devices is challenging due to the high dimensional hidden layers, e.g. a single weight matrix of a fully connected layer can have millions of parameters. By approximating this weight matrix by an MPS with small bond dimension, the authors in [Nov+15] were able to compress state-of-art image detection networks by a factor of seven with only minor performance penalties. Other examples of applications of tensor decompositions in deep learning include regularization [Tai+15] and deep representation learning [YH16].

Another use of the MPS tensor format in machine learning was pioneered in [MS16]. They consider a binary support vector machine classifier with a decision function

$$f(X) = \langle W, X \rangle \,. \tag{4.28}$$

Here, X is the input vector and W is the model vector. The goal is to determine W in such a way that $f(X) > 0$ if X is in the "yes" class and $f(X) < 0$ otherwise. By using an MPS representation for W and X, they are able to efficiently learn such a model even in high-dimensional settings. A similar approach can also be applied to learning polynomial classifiers [Che+17] as well as to image compression and classification [BPT15; Ben+17].

4.2. The Python Library mpnum

The Python library MPNUM [Sue17a] was developed during this work to simplify the process of prototyping numerical algorithms with MPS. Its main design principles are flexibility, user-friendliness, and expandability. We placed special emphasis on making the API of MPNUM as accessible as possible for researchers with little background in programming: Compared to the existing TT-toolboxes for Matlab and Python [Ose18a; Ose18b], we used object-oriented design principles to make the syntax as close to mathematical notation as possible and to hide the details of the MPS representation for the user. We were partly inspired by the fantastic ITensor library [Sto18] for C++, but preferred an implementation in pure Python due to its ease of use and widespread use in science. Beside our own work, MPNUM has been used in the study of experimental techniques in quantum optics [Sch+17a; Sch+17b], optical quantum simulation [Dha+18], as well as quantum many-body tomography [Lan+17]

4.2.1. The MPArray class

In this section, we exemplify the usage of mpnum in the context of quantum physics. The main goal is not to provide a comprehensive introduction, but to showcase the main design choices and goals of mpnum: flexibility, user-friendliness, and expandability. For a more thorough reference, we refer the reader to the online documentation under `http://mpnum.readthedocs.io/en/latest/`. Let us start by importing the necessary packages.

```
>>> import numpy as np
... import mpnum as mp
```

The fundamental data structure of mpnum is the `MPArray`, which stands for *matrix product array*. It is composed of an arbitrary number of local tensors with an arbitrary number of legs per site arranged in a linear chain. MPS and MPOs are special cases of this structure with one and two legs per site, respectively.

We start by performing the TT-SVD from Eq. (4.10) on a random tensor $X \in \mathbb{R}^{2^{10}}$.

```
>>> shape = 10 * (2,)
... X = np.random.randn(*shape)
... X /= np.linalg.norm(X.ravel())
... X_mps = mp.MPArray.from_array(X, ndims=1)
... X_mps.ndims

(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
```

This computes the MPS representation of `X`. By specifying `ndim=1`, we make sure the resulting tensor has one leg per site, which we check by `X_mps.ndims`.

Note here and throughout the rest of this section that the internal representation of the local tensors as a list of `numpy.ndarray` are hidden from the user behind an acccessible, high-level interface. However, direct access to the local tensors is provided using the `MPArray.lt` property, e.g. to compute how many floating point numbers are used in the MPS representation:

```
>>> sum(M.size for M in X_mps.lt)

2728
```

This is more than twice as large as the number of components for `X` itself, which is 2^{10}, or

```
>>> X.size
```

```
1024
```

Since the original tensor X is generated by sampling its components from a normal distribution, it is not compressible in MPS form. We see that the ranks of the tensor are exponentially increasing towards the middle as expected from Eq. (4.14)

```
>>> X_mps.ranks

(2, 4, 8, 16, 32, 16, 8, 4, 2)
```

Furthermore, even a moderate compression incurs a large approximation error.

```
>>> X_compressed, overlap = X_mps.compression(rank=11)
... overlap

0.6676040905553423

>>> X_compressed.ranks

(2, 4, 8, 11, 11, 11, 8, 4, 2)
```

Let us now demonnstrate the `MPArray` class for a compressible state, e.g. the W-state.

```
>>> from qutip.states import w_state
...
... psi = w_state(10).data.toarray().reshape((2,) * 10)
... psi_mps = mp.MPArray.from_array(psi, ndims=1)
... overlap = psi_mps.compress(rank=2)
...
... overlap

0.9999999999999996

>>> psi_mps.ranks

(2, 2, 2, 2, 2, 2, 2, 2, 2)
```

Note that in contrast to the previous case, we use in-place compression to reduce memory consumption. Clearly, the rank 2 MPS approximates the W-state up to numerical precission and requires staggeringly fewer parameters.

```
>>> sum(M.size for M in psi_mps.lt)
```

```
72
```

One main motivation behind encapsulating the local tensors in the `MPArray` data type is to ensure that they represent a valid MPS at all times and prevent common errors such as mismatch of the virtual dimensions. Furthermore, it allows us to keep track of the canonical form of the tensor.

```
>>> psi_mps.canonical_form

(9, 10)
```

Here, the first number indicates the index up to which all local tensors are left-normalized and the second number the index after which all local tensors are right-normalized. The `psi_mps` tensor in this example is in left-canonical form according to Definition 4.3. If we change one of the local tensors, e.g. by rescaling, the canonical form changes.

```
>>> M = psi_mps.lt[3]
... psi_mps.lt.update(3, 2 * M)
... psi_mps.canonical_form

(3, 10)
```

To bring it back to full canonical form, we need to call the appropriate method.

```
>>> psi_mps.canonicalize(left=9)
... psi_mps.canonical_form

(9, 10)
```

4.2.2. Arithmetic Operations

We now demonstrate the high-level interface for arithmetic operations on `MPArray` by simulating the preparation of a N-qubit GHZ state

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}} \left(|0, \ldots, 0\rangle + |1, \ldots, 1\rangle\right) \tag{4.29}$$

One possible circuit for this task is a succesive application of CNOT gates~[NC10]

$$|\mathrm{GHZ}\rangle = \mathrm{CNOT}_{N-1,N} \ldots \mathrm{CNOT}_{1,2} H_1 |0, \ldots, 0\rangle. \tag{4.30}$$

Here H_i is a Hadamard gate on the i-th qubit and $\mathrm{CNOT}_{i,j}$ denotes a controlled not gate with control on qubit i and target on qubit j. We start by defining the necessary local operations.

```
>>> from qutip.qip.gates import hadamard_transform
... from qutip.qip.gates import cnot as cnot_transform
...
... hadamard_local = hadamard_transform().data.toarray()
... cnot_local = cnot_transform().data.toarray()
... cnot_local

array([[1.+0.j, 0.+0.j, 0.+0.j, 0.+0.j],
       [0.+0.j, 1.+0.j, 0.+0.j, 0.+0.j],
       [0.+0.j, 0.+0.j, 0.+0.j, 1.+0.j],
       [0.+0.j, 0.+0.j, 1.+0.j, 0.+0.j]])
```

Next, generate the initial state in MPS form and convert the local operators to the `MPArray` data type.

```
>>> N = 10
...
... ket_down_local = np.array([1, 0], dtype=complex)
... ket_down = mp.MPArray.from_kron(N * [ket_down_local])
... hadamard = mp.MPArray.from_array(hadamard_local, ndims=2)
... len(ket_down)

10

>>> len(hadamard)

1
```

Note that the initial state $|0, \ldots, 0\rangle$ is a product, and hence, can be represented by an MPS of rank 1.

```
>>> ket_down.ranks

(1, 1, 1, 1, 1, 1, 1, 1, 1)
```

Since `cnot_local` is in matrix form, we cannot directly perform the TT-SVD on it. First, we have to convert it to a tensor of order four -- two legs per site -- and rearrange the legs in such a way such that legs from the same site are adjacent.

```
>>> cnot_local = cnot_local.reshape(4 * (2,)).transpose((0, 2, 1, 3))
... cnot =  mp.MPArray.from_array(cnot_local, ndims=2)
... len(cnot)
```

```
2

>>> cnot.ranks

(4,)
```

Now we can start to perform the circuit from Eq. (4.30).

```
>>> ket_ghz = mp.partialdot(hadamard, ket_down, start_at=0)
... ket_ghz.ranks

(1, 1, 1, 1, 1, 1, 1, 1, 1)
```

The function `partialdot` performs an efficient contraction of two `MPArray` of possibly unequal length. The result is an `MPArray` of the same order and -- since `hadamard` is a one-body operator -- of the same rank. We continue by applying the first CNOT, which results in an MPS with higher rank as CNOT entangles the two qubits on site one and two.

```
>>> ket_ghz = mp.partialdot(cnot, ket_ghz, start_at=0)
... ket_ghz.ranks

(4, 1, 1, 1, 1, 1, 1, 1, 1)
```

The other CNOT gates follow similarly.

```
>>> for site in range(1, N - 1):
...     ket_ghz = mp.partialdot(cnot, ket_ghz, start_at=site)
... ket_ghz.ranks

(4, 4, 4, 4, 4, 4, 4, 4, 4)
```

This yields an GHZ state in MPS representation.

A different approach is to simply generate the GHZ-state~(4.29) as a diagonal tensor. The two tensors are equal up to numerical precission.

```
>>> ket_ghz2 = mp.diagonal_mpa(np.array([1, 1]), N)
... ket_ghz2 /= mp.norm(ket_ghz2)
... mp.norm(ket_ghz - ket_ghz2)

1.5700924586837752e-16
```

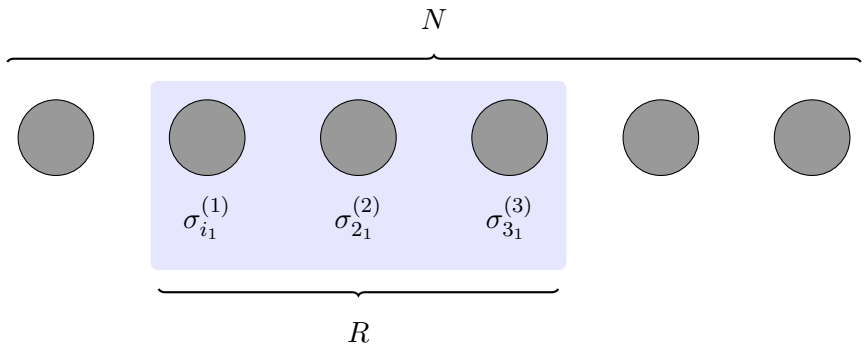


Figure 4.5.: The local measurements used for the reconstruction of MPS, MPO, and unitary channels in [Cra+10; Bau+13a; Bau+13b; Lan+17; Hol+15]. These consist of informationally complete measurements on blocks of R consecutive qudits, e.g. all Pauli product measurements on R qudits.

4.3. Efficient low-rank tensor reconstruction

So far, we have seen that the MPS tensor format provides the means to efficiently represent and manipulate tensors, which occur in practical applications in e.g. quantum physics or machine learning. In this context, the question arises: How can we *efficiently* recover this representation in practice from measurable quantities? Here, "efficiency" refers to two different aspects. On the one hand, we want to bound the number of measurements required to perform reconstruction, i.e. the *sample complexity*. On the other hand, it refers to the *computational complexity* as outlined in Section 2.2. Both aspects are especially critical for tensor reconstruction as naïve approaches generally suffer from the curse of dimensionality.

We first review existing work on efficient QSE using the MPS tensor format, which relies on local measurements on few neighboring sites, in Section 4.3.1 Although the notion of "tracing out" unobserved degrees of freedom is natural in quantum mechanics, it does not exist for the linear measurement model considered in this work, which we introduce also in said section. In Section 4.3.2, we describe the alternating least squares algorithm used for reconstruction from general product measurements, which is further analysed in Section 4.3.3. In Section 4.3.4 we consider the special case of Gaussian product measurements and present numerical reconstruction experiments in Section 4.3.5.

4.3.1. Existing work

In the context of quantum estimation, existing work addresses this question for the reconstruction of MPS [Cra+10], MPDO [Bau+13a; Bau+13b; Lan+17], and unitary

quantum processes [Hol+15] from local measurements. More precisely, they consider informationally complete measurement on blocks of length R as depicted in Fig. 4.5. Since there are exactly $N-R+1$ such blocks, this requires a number of measurements scaling as $\mathcal{O}(N \times d^R)$ compared to the $\mathcal{O}(d^N)$ scaling of full-fledged quantum state tomography. Numerical experiments in [Cra+10; Bau+13a; Bau+13b] demonstrate successful recovery of W-states as well as ground and thermal states of nearest-neighbor Hamiltonians for small values of R independent of N. Naturally, these are all examples of MPS or MPO with small bond dimension. The drastically reduced sampling complexity makes the approaches efficient and viable for large-scale quantum experiments. However, only the numerically inferior algorithm[5] from [Bau+13a] comes with a proof of convergence. Similar rigorous recovery guarantees have only been proven for comparable, but inefficient versions of the algorithms in [Cra+10; Bau+13b].

The question that motivated the work presented in this chapter is whether we can identify *other* measurement schemes and algorithms that *provably* allow for efficiently reconstructing low-rank tensors. However, a rigorous analysis of the local measurement model introduced above is very challenging. Furthermore, it requires "tracing out" the $N - R$ unobserved sites for each measurements, which relies on the operator-structure of mixed quantum states. For a general linear measurement model on tensors, there is no natural notion of partial trace, and hence, no R-local measurements. Therefore, we consider the problem of efficient reconstruction from a linear measurement model with product tensors here.

Recall that we already encountered a similar question in Chapter 3 for the problem of *low-rank matrix recovery*. The latter is concerned with the question, under which conditions we can recover a low-rank matrix $X \in \mathbb{C}^{d\times d}$ from m linear measurements of the form[6]

$$b^{(l)} = \langle A^{(l)}, X\rangle\,, \quad l = 1, \dots, m. \tag{4.31}$$

Here, $A^{(l)} \in \mathbb{C}^{d\times d}$ denote the measurement matrices and

$$\langle A, X\rangle = \operatorname{tr} A^\dagger X \tag{4.32}$$

the Frobenius inner product. For a general matrix X, the number of measurements m needs to scale as d^2 by simple parameter counting. However, by exploiting the low-rank structure of X, we can reduce m. For example, Theorem 3.5 shows that we can recover any positive semi-definite rank-1 matrix X from only $m = \mathcal{O}(d)$ measurements provided the $A^{(l)}$ are sampled from an appropriate distribution of

[5] T. Baumgratz, private communications.

[6] Note that we take up the notation from Chapter 3 for the rest of this chapter, where the index l in $A^{(l)}$ labels different measurement tensors, and not their local tensors.

random matrices. Furthermore, said theorem also provides an efficient reconstruction algorithm, namely the semi-definite program (3.12).

More generally, we can recover any $X \in \mathbb{C}^{d\times d}$ with $\operatorname{rank} X = r$ from $m = \mathcal{O}(dr)$ randomly chosen measurements [CP11; KRT17]. Intuitively, this sample complexity is asymptotically optimal since we need at least $2dr$ complex parameters to specify the left- and right-singular vectors of X, see [ENP12; LLB16] for rigorous lower bounds.

Here, we consider the generalization of low-rank matrix recovery to higher order tensors[7] $X \in \mathbb{R}^{d^N}$ of low MPS-rank. For this purpose, we study an *Alternating Least Squares* (ALS) algorithm. The observable quantities are – analogous to Eq. (4.31) – given by overlaps with measurement tensors $A^{(l)} \in \mathbb{R}$

$$b^{(l)} = \langle A^{(l)}, X\rangle \tag{4.33}$$

with the Frobenius inner product of tensors defined by Eq. (4.21).

The field of low-rank tensor recovery has attracted increasing attention in recent years. Especially the problem of tensor completion, i.e. inferring missing values of a low-rank tensor, has been thoroughly studied due to its broad applicability in computer vision, neuroscience, remote sensing, and context-aware recommender systems [LL10; ZH16; WNH14]. In contrast to our work, most analytical results in this area are concerned with different notions of tensor rank such as the Tucker rank [KSV14; Zha16] or the canonical tensor rank [KS13; PS17; GPY17]. The Tucker model still requires exponentially many parameters, and therefore, is unsuitable for our purposes. Although the canonical tensor representation captures the structure of many applications very well, it has some drawbacks in practice [KB09]: Approximation with fixed canonical rank in the Frobenius norm can be ill-posed [DL08]. Also, no equivalently tight bound on the approximation error as Eq. (4.22) is known for the canonical format.

The best analytical recovery guarantees for low-MPS rank tensor completion require – to the best of the authors knowledge – a number of measurements scaling exponentially in the order of the tensor [Phi+16]. Although their result gives a square-root advantage compared to naïve reconstruction, it is still infeasible for high-order tensors. However, numerical investigations using an alternating least squares algorithm similar to the one used in this work demonstrate successful recovery with sub-exponential sample complexity [GKK15a; WAA16].

Stronger analytical results exist for a different measurement ensemble: The authors in [RSS15; RSS17], consider fully Gaussian measurements, i.e. measurement tensors with independent components sampled from a normal distribution. In this approach, a number of measurements scaling polynomially in the crucial parameters d, N,

[7] In contrast to the rest of this chapter, we consider real tensors in order to simplify notation.

and r is sufficient to recover any rank r tensor. Although their result is highly efficient in the sample complexity, it is still infeasible for high-order tensors due to the exponentially scaling memory requirement for the measurement tensors and the resulting exponential runtime of any algorithm using them.

In existing work [HRS12; RU13], the authors proof local convergence result for general minimization problems in the MPS format. Their work shows that alternating minimization-type algorithms can be used to provable find local minima of a large class of cost functions.

In conclusion, the existing rigorous work on recovering low MPS-rank tensors is currently situated at two different ends of a spectrum: On the on hand, there is tensor completion, which is highly relevant for practical applications and efficiently implementable, but without any recovery guarantees for a polynomially scaling number of measurements. On the other hand, the authors in [RSS15] prove reconstruction with near-optimal sample complexity, but their full Gaussian measurements are computationally very demanding.

This situation can be understood better by comparing to the history of low-rank matrix recovery and compressive sensing – the related problem of reconstructing sparse vectors from few measurements. The first rigorous on compressed sensing from Gaussian measurements were due to Donoho [Don06] in 2006. In the same year, Candes, Romberg, and Tao – a Fields Medalist – proved guarantees for orthonormal basis vector measurements [CRT06], which can be considered as the analogue of matrix completion for vectors. Candes and Tao also proved the first matrix recovery guarantees for matrix completion [CT10], while the first guarantees for Gaussian matrices were proven in [RFP10]. Here, we are going to study a measurement ensemble for tensors that combines the advantages of both approaches. We take the measurement tensors to be of the form

$$A^{(l)} = a_1^{(l)} \otimes \cdots a_N^{(l)} \tag{4.34}$$

with local tensors randomly chosen from a standard multivariate normal distribution $a_k^{(l)} \sim \mathcal{N}(0, \mathbb{1}_d)$. In contrast to the fully Gaussian model from [RSS15; RSS17], the measurement tensors (4.34) can be represented efficiently as MPS of unit rank. And, unlike the tensor completion problem, history does not suggest that its solution requires at least one Fields Medalist.

The rest of this section is structured as follows: In Section 4.3.2, we introduce the *alternating least squares* (ALS) algorithm for tensors, which we analyse analytically in Section 4.3.3 for general rank-1 measurements. This section also contains the main analytical result of this section, namely a sufficient condition for recovery, which connects the deviation of the initialization from the true value and inherent properties of the measurements. The particular case of Gaussian rank-1 Gaussian measurements is further investigated in Section 4.3.4. As the analysis of Gaussian

product measurements poses tremendous challenges, this section contains work in progress. We develop mathematical tools, which might be useful in future work, and perform detailed numerical experiments investigation of the properties of said measurements. Finally, in Section 4.3.5, we report on numerical simulations of tensor recovery via ALS and some optimization strategies for scaling the implementation to large tensors.

4.3.2. The alternating least squares algorithm

To recapitulate, the goal is to recover a tensor $X \in \mathbb{R}^{d^N}$ with MPS-rank[8] r from M linear measurements of the form

$$b^{(l)} = \langle A^{(l)}, X \rangle \,, \ldots \, l = 1, \ldots, M, \tag{4.35}$$

where $A^{(l)}$ is a product of Gaussian vectors as defined in Eq. (4.34). For this purpose, we want an efficient reconstruction algorithm and M to scale polynomially in the parameters d, N, and r.

The idea to recover X is simply to find the tensor Y of desired rank that minimizes the empirical ℓ_2 error

$$\frac{1}{M} \sum_l \left(b^{(l)} - \langle A^{(l)}, Y \rangle \right)^2 . \tag{4.36}$$

In general, this problem is hard to solve directly since the space of MPS of given rank is non-convex. Nevertheless, in the case of $M = \mathcal{O}(((N-1)r^3 + dNr) \log dr)$ fully Gaussian measurements, a projected gradient descent on (4.36) is able to recover X [RSS15; RSS17]. However, the same proof techniques are not suitable for the Gaussian rank-one measurements under consideration here. This is due to the fact that the measurement tensors lie in the variational class of tensors we try to recover. For more details, see Appendix B in [ZJD15].

Instead of updating all local tensor simultaneously as in the gradient descent, we iteratively optimize the empirical error over a single local tensor at a time. Since the minimization of Eq. (4.36) over a single local tensor is a linear-least squares problem,

[8] From now on, we simply say "rank" instead of "MPS-rank".

it can be solved efficiently. The resulting algorithm is presented below:

Algorithm 1: Alternating Least Squares (ALS) for ℓ_2 minimization

Input : Number of epochs H and initialization MPS X_{init} of order N, rank r, and local dimension d, measurement tensors $A^{(l)} = a_1^{(l)} \otimes \cdots \otimes a_N^{(l)}$ and measurement outcomes $b^{(l)}$ with $i = 1, \ldots, H \times N \times m$ divided into HN batches of size m, which are denoted by $\left(A^{(h,n;l)}, b^{(h,n;l)}\right)_l$ $(h \in [H], n \in [N])$

1 $Y \leftarrow X_{\text{init}}$
2 **for** $h \leftarrow 1$ **to** H **do**
  `/* right-normalize all local tensors, i.e. bring` Y `to right-canonical form */`
3 `right_canonicalize`(Y)
4 **for** $n \leftarrow 1$ **to** N **do**
5 **for** $l \leftarrow 1$ **to** m **do**
  `/* contract` $A^{(h,n;l)}$ `with all but n-th local tensors */`
6 $B^{(l)} \leftarrow$ `contract`$(A^{(h,n;l)}, Y_{[N]\setminus n})$
7 $\hat{Z} \leftarrow \operatorname{argmin}_Z \sum_l \left(b^{(h,n;l)} - B^{(l)} Z\right)^2$
  `/* update the n-th local tensor inplace with a left-normalized form of` $\hat{Z}$ `*/`
8 $Y_n \leftarrow$ `left_normalize`$(\hat{Z})$

Output: Y

In this version of the alternating least squares scheme, we start updating the left-most tensor and then move through the chain all the way to the right. When we reach the last tensor of the MPS, we start again on the left after bringing the MPS to right-canonical form. Hence, all tensors to the left and right of the currently updated tensor are left- and right-normalized, respectively. This process for a total of H epochs.

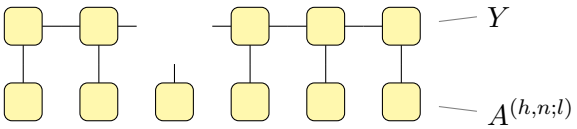
The crucial step in this algorithm is the local minimization in line 7. It amounts to keeping all but the n-th local tensor of Y fixed and minimizing the empirical error over Y_n. The minimizer $\hat{Z}$ can be computed efficiently from the linear least squares problem

$$\hat{Z} = \operatorname{argmin}_Z \sum_l \left(b^{(h,n;l)} - B^{(l)} Z\right)^2 = \|b^{(h,n)} - BZ\|_{\ell_2}^2. \tag{4.37}$$

Here, the rows of the matrix[9] $B \in \mathbb{R}^{m \times (dr^2)}$ are defined in line 6 in terms of partial

[9] More specifically, we have $B \in \mathbb{R}^{m \times (rd)}$ for $n = 1$ or $n = N$.

contractions of $A^{(h,n;l)}$ and Y leaving out the n-th local tensor of Y:

$B^{(l)} =$ 

(4.38)

Therefore, the solution $\hat{Z}$ of Eq. (4.37) is a dr^2 dimensional vector, which can be reshaped to the correct form. If we replace the exact local minimization in Eq. (4.37) by a finite gradient descent step, we obtain a standard nonlinear block Gauss–Seidel iteration [Sch62].

Finally, note that the sample splitting into H epochs at the beginning is necessary for the analysis of the algorithm below as it requires stochastically independent updates for each micro-iteration. Therefore, we use a fresh batch of measurement tensors and measurement values in each step, which results in a total number of $M = HNm$ measurements. Provided that both, the number of epochs H and the batch size m scale polynomially in the system's parameters, then so does M. However, numerical experiments in Section 4.3.5 show that this resampling is unnecessary in practice.

4.3.3. Analysis of the ALS

Alternating algorithms such as ALS updating only a few local tensors at a time are a very common approximation technique for circumventing intractabilities when dealing with MPS. Well known examples include variational compression and DMRG [Sch11] – an iterative algorithm for approximating the smallest eigenvalue of a hermitian MPO. Local convergence of these alternating algorithms has been proven for a large class of problems in [HRS12; RU13]. These results show that alternating minimization algorithms of many cost functions converge to a local minimum. However, it remains to be shown that the minimizer of the empirical ℓ_2 error (4.36) with given rank is equal to X, i.e. that the given measurements suffice to identify X.

For this purpose, we generalize the ideas from the matrix case [ZJD15] to the tensor case. More precisely, We adapt the idea from [ZJD15] to analyse the micro-iterations of Alg. 1 directly by deriving a closed form expression for the minimizer of Eq. (4.37). However, as the notation and the analysis is much more involved in the tensor case, we only treat the case of a product signal tensor, i.e. X is assumed to have rank one. Furthermore, we also assume w.l.o.g. that X is normalized in Frobenius norm, i.e. $\|X\|_2 = 1$. Then, X can be written as a tensor product of N normalized vectors $x_i \in \mathbb{R}^d$ and we have

$$b^{(l)} = \langle A^{(l)}, X \rangle = \prod_{i=1}^{N} \langle a_i^{(l)}, x_i \rangle \,. \tag{4.39}$$

One main result of this chapter is Theorem 4.7. It shows that under certain assumptions on the initialization and measurement tensors, each micro-iteration brings the local tensor closer to its true value in a suitable metric. In the case of rank-1 tensor reconstruction, this metric between the local tensors x and y with $\|x\|_{\ell_2} = \|y\|_{\ell_2} = 1$ is given by

$$\operatorname{dist}(x, y) = \sqrt{1 - |\langle x, y\rangle|^2}. \tag{4.40}$$

It is known as *infidelity* if $x, y \in \mathbb{C}^d$ represent pure quantum states [NC10]. Equation (4.40) is also the one-dimensional special case of the *principle angle distance* of subspaces [GV12], which is the appropriate distance measure for higher-rank generalizations of the results presented here. The following lemma relates the "global" distance of two tensors, which is measured in Frobenius norm, to the distances of the local tensors measured by (4.40). Therefore, it shows how a reduction of the errors of the local tensors leads to a reduction of the error of the full tensor measured in Frobenius norm.

Lemma 4.4. *Let $X = \otimes_{i=1}^N x_i$ and $Y = \otimes_{i=1}^N y_i$ be two product tensors of unit norm. Assume that for each $i \in [N]$, we have $\|x_i\|_{\ell_2} = \|y_i\|_{\ell_2} = 1$ and*

$$1 - \langle x_i, y_i\rangle^2 \leq \delta_i^2. \tag{4.41}$$

Then,

$$\min_{\eta=\pm 1} \|X - \eta Y\|_2 \leq \sqrt{2} \sum_{i=1}^N \delta_i. \tag{4.42}$$

Proof. First, note that Eq. (4.41) implies

$$\min_{\eta=\pm 1} \|x_i - \eta y_i\|_{\ell_2}^2 = \min_{\eta=\pm 1} 2(1 - \eta \langle x_i, y_i\rangle) = 2(1 - |\langle x_i, y_i\rangle|) \tag{4.43}$$

$$\leq 2(1 - \langle x_i, y_i\rangle^2) \leq 2\delta_i^2. \tag{4.44}$$

Therefore, we get for the Frobenius norm distance of the full tensors

$$\min_{\eta=\pm 1} \|X - \eta Y\|_2 = \min_{\eta\in\{\pm 1\}} \|x_1 \otimes \cdots \otimes x_N - \eta\, y_1 \otimes \cdots y_N\|_2 \tag{4.45}$$

$$\leq \min_{\eta,\xi\in\{\pm 1\}} (\|x_1 \otimes x_2 \otimes \cdots \otimes x_N - \eta\xi\, x_1 \otimes y_2 \otimes \cdots \otimes y_N\|_2 + \|\eta\xi\, x_1 \otimes y_2 \otimes \cdots \otimes y_N - \eta\, y_1 \otimes y_2 \otimes \cdots y_N\|_2) \tag{4.46}$$

$$= \min_{\eta\in\{\pm 1\}} \|x_2 \otimes \cdots \otimes x_N - \eta\, y_2 \otimes \cdots \otimes y_N\|_2 + \min_{\eta\in\{\pm 1\}} \|x_1 - \eta\, y_1\|_2. \tag{4.47}$$

By telescoping this argument, we get

$$\min_{\eta=\pm 1} \|x_i - \eta y_i\|_{\ell_2}^2 \leq \sum_{i=1}^{N} \left(\min_{\eta\in\{\pm 1\}} \|x_i - \eta y_i\|_2 \right) \leq \sqrt{2} \sum_{i=1}^{N} \delta_i, \tag{4.48}$$

which completes the proof. □

We now turn to the main problem in this section, namely investigating the ALS update step. In the following, we denote by $[N] = 1, \ldots, N$ the set of all integers up to N and by $[N]_{\setminus j} = [N] \setminus \{j\}$. For our analysis, the following two conditions on the measurement tensors $A^{(l)}$ are crucial:

1. **Concentration of the B_j operators**: For all $j \in [N]$, let $v_i \in \mathbb{R}^d$ with $\|v_i\|_2 = 1$ ($i \in [N]_{\setminus j}$) independent of all the $a_i^{(l)}$. Define

$$B_j\left((v_i)_{i\in[N]_{\setminus j}}\right) = \frac{1}{m}\sum_{l=1}^{m} \left(\prod_{i\neq j} \left\langle a_i^{(l)}, v_i \right\rangle \right)^2 \left| a_j^{(l)} \right\rangle \left\langle a_j^{(l)} \right|, \tag{4.49}$$

   Then, the smallest eigenvalue of $B_j\left((v_i)_{i\in[N]_{\setminus j}}\right)$ should satisfy

$$\lambda_{\min}\left(B_j\left((v_i)_{i\in[N]_{\setminus j}}\right)\right) \geq \delta_{\mathrm{B}} \tag{4.50}$$

   for some constant $\delta_{\mathrm{B}} > 0$

2. **Concentration of the G_j operators**: For all $j \in [N]$, let $v_i, v_i^{\perp} \in \mathbb{R}^d$ ($i \in [N]_{\setminus j}$) independent of all the $a_i^{(l)}$. Furthermore, they should satisfy $\|v_i\|_2 = \|v_i^{\perp}\| = 1$ and $\langle v_i, v_i^{\perp}\rangle = 0$ for $i \in [N]_{\setminus j}$. Define

$$G_j\left((v_i, v_i^{\perp})_{i\in[N]_{\setminus j}}\right) = \frac{1}{m}\sum_{l=1}^{m} \left(\prod_{i\neq j} \left\langle a_i^{(l)}, v_i \right\rangle \left\langle a_i^{(l)}, v_i^{\perp} \right\rangle \right) \left| a_j^{(l)} \right\rangle \left\langle a_j^{(l)} \right|. \tag{4.51}$$

   Then, the largest singular value of $G_j\left((v_i, v_i^{\perp})_{i\in[N]_{\setminus j}}\right)$ should satisfy

$$\left\| G_j\left((v_i, v_i^{\perp})_{i\in[N]_{\setminus j}}\right) \right\|_{2\to 2} \leq \delta_{\mathrm{G}}. \tag{4.52}$$

   for some constant $\delta_{\mathrm{G}} > 0$

In Section 4.3.4 we elaborate on work in progress with the goal of proving that random local tensors $a_i^{(l)}$ sampled independently from a standard d-variate Gaussian

distribution satisfy these conditions with high probability.

We start the analysis by examining a single ALS update step for the leftmost local tensor with $n = 1$. Denote by $\tilde{y}_1^{h+1}$ the minimizer $\hat{Z}$ in line 7 of Alg. 1 and by y_i^h the remaining local tensors of Y, which are kept fixed during this micro-iteration. The empirical error as a function of $\tilde{y}_n^{h+1}$ then reads

$$F\left(\tilde{y}_1^{h+1}\right) = \sum_l \left(\prod_i \langle a_i^{(l)}, x_i \rangle - \langle a_1^{(l)}, \tilde{y}_1^{h+1} \rangle \prod_{i \neq 1} \langle a^{(l)}, y_i^h \rangle \right)^2 \tag{4.53}$$

Since in the following, we only consider the optimization over $\tilde{y}_1^{h+1}$ and keep the other y_i^h fixed, we write $\tilde{y}_1$ and y_i (for $i > 1$), respectively, if there is no risk of confusion. In contrast to the $(y_i)_{i>1}$, $\tilde{y}_1$ is not normalized. We now derive an explicit representation of the minimizer of Eq. (4.53).

Lemma 4.5. *Denote by F the empirical ℓ_2 error defined in Eq.* (4.53)*. Then, the extremal point $\tilde{y}_1$ with $\frac{\partial F(\tilde{y}_1)}{\partial \tilde{y}_1} = 0$ is given by*

$$\tilde{y}_1 = \left(\prod_{i \neq 1} \langle x_i, y_i \rangle \right) x_1 - \tilde{B}_1^{-1} \tilde{G}_1 x_1 \tag{4.54}$$

with $\tilde{B}_1$ and $\tilde{G}_1$ given by

$$\tilde{B}_1 = \frac{1}{m} \sum_l \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right)^2 \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \tag{4.55}$$

$$\tilde{G}_1 = \frac{1}{m} \sum_l \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, \tilde{y}_i^{\perp} \right\rangle \left\langle a_i^{(l)}, y_i \right\rangle \right) \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \tag{4.56}$$

$$\tilde{y}_i^{\perp} = \left(|y_i\rangle\langle y_i| - \mathbb{1} \right) x_i \quad (i \in [N]_{\setminus 1}). \tag{4.57}$$

Proof. We begin by computing the derivative of F

$$\frac{\partial F(\tilde{y}_1)}{\partial \tilde{y}_1} = -2 \sum_l \left(\prod_i \left\langle a_i^{(l)}, x_i \right\rangle - \left\langle a_1^{(l)}, \tilde{y}_1 \right\rangle \prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right) \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right) a_1^{(l)}, \tag{4.58}$$

which is equal to zero if and only if

$$\sum_l \left(\left\langle a_1^{(l)}, \tilde{y}_1 \right\rangle \prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right) a_1^{(l)} = \sum_l \left(\prod_i \left\langle a_i^{(l)}, x_i \right\rangle \prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right) a_1^{(l)}. \tag{4.59}$$

Reordering terms and factoring out the terms with $i = 1$ gives

$$\sum_l \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, y_i \right\rangle \right)^2 a_1^{(l)} \left\langle a_1^{(l)}, \tilde{y}_1 \right\rangle = \sum_l \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, x_i \right\rangle \left\langle a_i^{(l)}, y_i \right\rangle \right) a_1^{(l)} \left\langle a_1^{(l)}, x_1 \right\rangle, \tag{4.60}$$

and therefore,

$$\tilde{B}_1 \left| \tilde{y}_1 \right\rangle = \underbrace{\frac{1}{m} \left[\sum_l \left(\prod_{i \neq 1} \left\langle a_i^{(l)}, x_i \right\rangle \left\langle a_i^{(l)}, y_i \right\rangle \right) \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \right]}_{=: \tilde{O}_1} x_1. \tag{4.61}$$

Note that $\tilde{B}_1 = B_1(y_2, \ldots, y_N)$ as defined in Eq. (4.49). Since we assume by Eq. (4.50) that the smallest eigenvalue of $\tilde{B}_1$ is larger than zero, it is invertible and we can multiply the last equation by its inverse and obtain

$$\tilde{y}_1 = \left(\prod_{i \neq 1} \langle x_i, y_i \rangle \right) x_i - \tilde{B}_1^{-1} \left(\left(\prod_{i \neq 1} \langle x_i, y_i \rangle \right) \tilde{B}_1 - \tilde{O}_1 \right) x_i, \tag{4.62}$$

where the first and second summand cancel each other. Finally, we simplify the expression in parentheses, which completes the proof:

$$\left(\prod_{i \neq 1} \langle x_i, y_i \rangle \right) \tilde{B}_1 - \tilde{O}_1 \tag{4.63}$$

$$= \frac{1}{m} \sum_l \left(\prod_{i \neq 1} \langle x_i, y_i \rangle \left\langle a_i^{(l)}, y_i \right\rangle^2 - \prod_{i \neq 1} \langle a_i^{(l)}, y_i \rangle \left\langle a_i^{(l)}, x_i \right\rangle \right) \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \tag{4.64}$$

$$= \frac{1}{m} \sum_l \left(\prod_{i \neq 1} \langle a_i^{(l)}, y_i \rangle \left(\langle x_i, y_i \rangle \left\langle a_i^{(l)}, y_i \right\rangle - \langle a_i^{(l)}, x_i \rangle \right) \right) \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \tag{4.65}$$

$$= \frac{1}{m} \sum_l \left(\prod_{i \neq 1} \langle a_i^{(l)}, y_i \rangle \left(\left\langle a_i^{(l)}, \langle y_i, x_i \rangle \, y_i - x_i \right\rangle \right) \right) \left| a_1^{(l)} \right\rangle \left\langle a_1^{(l)} \right| \tag{4.66}$$

$$= \tilde{G}_1. \tag{4.67}$$

□

From Eq. (4.54) we see that if the error term $\tilde{B}_1^{-1} \tilde{G}_1 x_1$ is small, then the overlaps of the remaining local tensors $\langle x_i, y_i \rangle$ for $i = 2, \ldots, N$ determine how close $\tilde{y}_1$ is to its true value x_1. The following lemma makes this precise.

Lemma 4.6. *Let* $y_1 := \frac{\tilde{y}_1}{\|\tilde{y}_1\|_{\ell_2}}$ *with* $\tilde{y}_1$ *given by Eq.* (4.54)*. Furthermore, define*

$$\kappa_1 = \frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}} \prod_{i \neq 1} |\langle x_i, y_i \rangle| - \prod_{i \neq 1} \sqrt{1 - \langle x_i, y_i \rangle^2}. \tag{4.68}$$

Assume that the conditions of Eqs. (4.50) *and* (4.52) *hold and* $\kappa_1 > 0$*, then we have*

$$1 - \langle x_1, y_1 \rangle^2 \leq \frac{\prod_{i \neq 1}(1 - \langle x_i, y_i \rangle^2)}{\kappa_1^2}. \tag{4.69}$$

Proof. First recall that $\tilde{B}_1 = B_1(y_2, \dots, y_N)$ as defined in Eq. (4.49). Since $\|y_i\|_{\ell_2} = 1$ and $\tilde{B}_1$ is positive definite, the lower bound on the smallest eigenvalue of $\tilde{B}_1$ in Eq. (4.50) translates to

$$\left\| \tilde{B}_1^{-1} \right\|_{2 \to 2} = \frac{1}{\lambda_{\min}(\tilde{B}_1)} \leq \frac{1}{\delta_{\mathrm{B}}}. \tag{4.70}$$

Furthermore, $\tilde{G}_1 = G_1(y_2, \tilde{y}_2^{\perp}, \dots, y_N, \tilde{y}_N^{\perp})$, and hence, normalization of the $\tilde{y}_i^{\perp}$ yields

$$\begin{aligned} \left\| \tilde{G}_1 \right\|_{2 \to 2} &\leq \delta_{\mathrm{G}} \prod_{i \neq 1} \|\tilde{y}_i^{\perp}\|_{\ell_2} \\ &\leq \delta_{\mathrm{G}} \prod_{i \neq 1} \sqrt{1 - \langle x_i, y_i \rangle^2}. \end{aligned} \tag{4.71}$$

Combining Eqs. (4.70) and (4.71) as well as the condition $\|x_1\|_{\ell_2} = 1$ yields the bound on the error term

$$\left\| \tilde{B}_1^{-1} \tilde{G}_1 x_1 \right\|_{\ell_2} \leq \frac{\delta_{\mathrm{G}}}{\delta_{\mathrm{B}}} \prod_{i \neq 1} \sqrt{1 - \langle x_i, y_i \rangle^2}. \tag{4.72}$$

Next, using Eq. (4.54), we can lower bound the overlap

$$|\langle x_1, \tilde{y}_1 \rangle| = \left| \prod_{i \neq 1} \langle x_i, y_i \rangle - \langle x_i, B_1^{-1} G_1 x_1 \rangle \right| \tag{4.73}$$

$$\geq \left| \prod_{i \neq 1} |\langle x_i, y_i \rangle| - \left| \langle x_1, B_1^{-1} G_1 x_1 \rangle \right| \right| \tag{4.74}$$

$$\geq \prod_{i \neq 1} |\langle x_i, y_i \rangle| - \left\| B_1^{-1} G_1 x_1 \right\|_{\ell_2} \tag{4.75}$$

$$\geq \prod_{i \neq 1} |\langle x_i, y_i \rangle| - \frac{\delta_{\mathrm{G}}}{\delta_{\mathrm{B}}} \prod_{i \neq 1} \sqrt{1 - \langle x_i, y_i \rangle^2}. \tag{4.76}$$

by the Cauchy-Schwartz inequality and Eq. (4.72). Along the same lines, we obtain for any $x_1^\perp$ with $\|x_1^\perp\|_{\ell_2} = 1$ and $\langle x_1, x_1^\perp\rangle = 0$

$$\begin{aligned}
\left|\langle x_1^\perp, \tilde{y}_1\rangle\right| &= \left|\left\langle x_1^\perp, x_1\right\rangle \prod_{i\neq 1}\langle x_i, y_i\rangle - \left\langle x_1^\perp, \tilde{B}_1^{-1}\tilde{G}_1 x_1\right\rangle\right| && (4.77)\\
&= \left|\left\langle x_1^\perp, \tilde{B}_1^{-1}\tilde{G}_1 x_1\right\rangle\right| && (4.78)\\
&\leq \left\|x_1^\perp\right\|_{\ell_2}\left\|\tilde{B}_1^{-1}\tilde{G}_1 x_1\right\|_{\ell_2} && (4.79)\\
&\leq \frac{\delta_{\mathrm{G}}}{\delta_{\mathrm{B}}}\prod_{i\neq 1}\sqrt{1-\langle x_i, y_i\rangle^2}. && (4.80)
\end{aligned}$$

Finally, note that by the Pythagorean Theorem, we have

$$\|\tilde{y}_1\|_{\ell_2}^2 = \langle x_1, \tilde{y}_1\rangle^2 + \langle x_1^\perp, \tilde{y}_1\rangle^2 . \tag{4.81}$$

Combining Eqs. (4.76), (4.80) and (4.81) then gives the final estimate

$$\begin{aligned}
1-\langle x_1, y_1\rangle^2 &= 1 - \frac{\langle x_1, \tilde{y}_1\rangle^2}{\|\tilde{y}_1\|_{\ell_2}^2} && (4.82)\\
&= \frac{\langle x_1^\perp, \tilde{y}_1\rangle^2}{\langle x_1, \tilde{y}_1\rangle^2 + \langle x_1^\perp, \tilde{y}_1\rangle^2} && (4.83)\\
&\leq \frac{\langle x_1^\perp, \tilde{y}_1\rangle^2}{\langle x_1, \tilde{y}_1\rangle^2} && (4.84)\\
&\leq \frac{\left(\frac{\delta_{\mathrm{G}}}{\delta_{\mathrm{B}}}\right)^2 \prod_{i\neq 1}\left(1-\langle x_i, y_i\rangle^2\right)}{\left(\prod_{i\neq 1}|\langle x_i, y_i\rangle| - \frac{\delta_{\mathrm{G}}}{\delta_{\mathrm{B}}}\prod_{i\neq 1}\sqrt{1-\langle x_i, y_i\rangle^2}\right)^2} && (4.85)
\end{aligned}$$

provided the right hand side of Eq. (4.76) is positive. This completes the proof. □

We are now ready to state the main theorem of this section. It shows that the Alternating Least Squares algorithm improves the overlap of each local tensor with its true value provided their initial values are close enough to the true value and the conditions (4.50) and (4.52) for the measurements hold.

Theorem 4.7. *Let $N > 2$. Assume that the initial values for the local tensors satisfy*

$$\left|\langle x_i, y_i^0\rangle\right| \geq \sqrt{1-\delta_{\mathrm{I}}^2} \quad 1 = 2, \ldots, N \tag{4.86}$$

for a constant $\delta_{\mathrm{I}} \geq 0$ *and the conditions* (4.50) *and* (4.52) *hold. Set*

$$\theta := \frac{\delta_{\mathrm{I}}^{N-2}}{\sqrt{\frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}}(1-\delta_{\mathrm{I}}^2)^{\frac{N-1}{2}} - \delta_{\mathrm{I}}^{N-1}}}. \tag{4.87}$$

If $\theta < 1$*, the local updates after completing one epoch satisfy*

$$\sqrt{1 - \langle x_i, y_i^1 \rangle^2} \leq \theta^{2^{i-1}} \delta_{\mathrm{I}}. \tag{4.88}$$

Proof. Note that the initialization condition (4.86) implies that

$$\kappa_1^0 = \frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}} \prod_{i \neq 1} \left| \langle x_i, y_i^0 \rangle \right| - \prod_{i \neq 1} \sqrt{1 - \langle x_i, y_i^0 \rangle^2} \tag{4.89}$$

$$\geq \frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}} \left(1 - \delta_{\mathrm{I}}^2\right)^{\frac{N-1}{2}} - \delta_{\mathrm{I}}^{N-1} \tag{4.90}$$

$$> 0, \tag{4.91}$$

since Eq. (4.87) would be ill-defined otherwise. Therefore, we have with Lemma 4.6

$$1 - \langle x_1, y_1^1 \rangle^2 \leq \frac{1}{\kappa_{\delta_{\mathrm{I}}}^2} \prod_{i \neq 1} (1 - \langle x_i, y_i^0 \rangle^2) \tag{4.92}$$

$$\leq \frac{\delta_{\mathrm{I}}^{2(N-2)}}{\kappa_{\delta_{\mathrm{I}}}^2} \delta_{\mathrm{I}}^2. \tag{4.93}$$

$$= \theta^2 \, \delta_{\mathrm{I}}^2. \tag{4.94}$$

Under the assumption that $\theta < 1$, Eq. (4.94) implies that the upper bound on the overlap of the first local tensors decreases. Similarly, we find for the second local tensor

$$1 - \langle x_i, y_2^1 \rangle^2 \leq \frac{1}{\kappa_{\delta_{\mathrm{I}}}^2} \left(1 - \left\langle x_1, y_1^1 \right\rangle^2\right) \prod_{i > 2} (1 - \left\langle x_i, y_i^0 \right\rangle^2) \tag{4.95}$$

$$\leq \frac{1}{\kappa_{\delta_{\mathrm{I}}}^2} \, \theta^2 \delta_{\mathrm{I}}^2 \, \delta_{\mathrm{I}}^{2(N-2)} \tag{4.96}$$

$$= \theta^4 \delta_{\mathrm{I}}^2. \tag{4.97}$$

Iterating this argument shows that for the i-th local tensor, we have

$$1 - \langle x_2, y_i^1 \rangle^2 = \theta^{2n_i} \delta_{\mathrm{I}}^2, \tag{4.98}$$

where n_i satisfies $n_1 = 1$ and $n_{i+1} = 1 + \sum_{j=1}^{i} n_j$, and hence, $n_i = 2^{i-1}$ □

Note that for $\theta = \text{const} < 1$, Eq. (4.88) implies an improbable fast, double-exponential convergence in N for some local tensors. Hence, we expect an N-dependent "constant" θ with $\theta \to 1$ as $N \to \infty$. Also, Eq. (4.88) suggests that the order in which we optimize the local tensors in Alg. 1 is not optimal: Due to the dependence of the right hand side of said equation on the site-index i, the rightmost local tensor benefits the most from the improvements of the previous steps. Therefore, an alternating left-right-left sweep may be more beneficial for improving the global error of the reconstruction as the latter is determined by the largest error of the local tensors according to Lemma 4.4. Such sweeping strategy also is widely used in related algorithms such as DMRG. In contrast to such alternating sweep strategies, one-directional sweeping requires additional canonicalization after each completed sweep in order to bring the tensor into the necessary form.

Let us now investigate the convergence condition on θ more closely. A straightforward computation shows that $\theta < 1$ is equivalent to

$$\frac{\delta_\mathrm{B}}{\delta_\mathrm{G}} > \left(\frac{\delta_\mathrm{I}^2}{1-\delta_\mathrm{I}^2}\right)^{\frac{N-1}{2}} \frac{1+\delta_\mathrm{I}}{\delta_\mathrm{I}}. \tag{4.99}$$

Hence, there is a trade-off between the measurement-constants δ_B and δ_G, and the initialization constant δ_I: On the one hand, if $\delta_\mathrm{I} > \frac{1}{\sqrt{2}}$, Eq. (4.99) the right hand side of Eq. (4.99) grows exponentially fast as a function of N. Since effectively, we need to lower bound $\frac{\delta_\mathrm{B}}{\delta_\mathrm{G}}$, Eq. (4.99) necessitates very strong concentration properties of the measurement ensemble. On the other hand, if $\delta_\mathrm{I} < \frac{1}{\sqrt{2}}$, the right hand side goes to zero exponentially fast in N. Therefore, the better the initialization, the more leeway we have for the bounds of the concentration constants δ_B and δ_G.

Finally, note that Theorem 4.7 only guarantees that the ALS algorithm improves the *absolute values* of the overlaps of the local tensors. Therefore, we are only able to reconstruct the local tensors up to a sign.[10] This is due to the Gauge symmetry of MPS mentioned in Section 4.1.2: Since we deal with a rank-1 MPS and fix all the norms of the local tensors, the remaining gauge freedom is exactly given by transformations of the form

$$x_i \mapsto -x_i \text{ and } x_j \mapsto -x_j \tag{4.100}$$

for $i \neq j$. The ability to reconstruct all local tensors only up to sign implies that we are able to reconstruct X only up to a global sign using ALS as we show in Lemma 4.4 below. However, in our idealized scenario without noise, the sign can be easily recovered simply by comparing $b^{(l)} = \langle A^{(l)}, X\rangle$ to $\langle A^{(l)}, Y^\sharp\rangle$ for any l provided $b^{(l)}$ is larger than the small remaining reconstruction error. Here, $Y^\sharp$ denotes the final output of ALS.

[10] Or up to a phase factor in the complex case.

4.3.4. Gaussian measurements

The results from Section 4.3.3 apply to general rank-1 measurements

$$A^{(l)} = a_1^{(l)} \otimes \cdots \otimes a_N^{(l)}, \tag{4.101}$$

which satisfy the fundamental concentration conditions in Eqs. (4.50) and (4.52). In this section, we consider measurements of the form (4.101), where the local tensors are chosen independently from a standard normal d-variate distribution, i.e. $a_N^{(l)} \sim \mathcal{N}(0, \mathbb{1}_d)$. In this case, the operators $B_j(v)$ defined in Eq. (4.49) are of the form

$$B_j(v) = \frac{1}{m} \sum_{l=1}^{m} Y_j^{(l)} \left| a_j^{(l)} \right\rangle \left\langle a_j^{(l)} \right|, \tag{4.102}$$

where

$$Y_j^{(l)} = \prod_{i \neq j} \left\langle a_i^{(l)}, v_i \right\rangle^2 = \prod_{i=1}^{N-1} {g_i^{(l)}}^2 \tag{4.103}$$

is a product of squares of $N-1$ independent standard Gaussians $g_i^{(l)} \sim \mathcal{N}(0,1)$. Similarly, for $G_j(v, v^\perp)$ defined in Eq. (4.51)

$$G_j(v, v^\perp) = \frac{1}{m} \sum_{l=1}^{m} X_j^{(l)} \left| a_j^{(l)} \right\rangle \left\langle a_j^{(l)} \right|, \tag{4.104}$$

with

$$X_j^{(l)} = \prod_{i=1}^{2(N-1)} g_i^{(l)} \tag{4.105}$$

being a product of $2(N-1)$ independent standard Gaussians. Therefore, the problem of proving the concentration properties Eqs. (4.50) and (4.52) requires a better understanding of the distribution of products of Gaussian random variables.

In this section we are going to summarize the results from [SSK17a; SSK17b], where we compute a power-log series expansion for the cumulative distribution function of products of independent standard Gaussians as well their absolute values and squares. As the main work in these publications was done by the first author, we only sketch the results here. In numerical simulations, we also show that truncations of said expansions at the first-order provide very close approximations, and therefore, may be used to derive strong concentration properties for said random variables. However, since at this time, we do not have explicit error bounds for these truncations, we investigate the concentration properties of the operators (4.102) and (4.104) numerically at the end of this section.

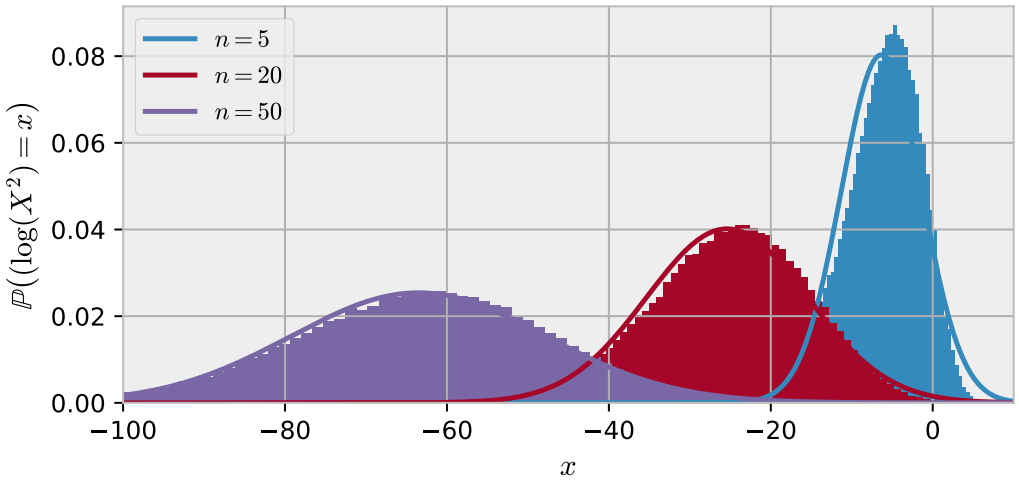


Figure 4.6.: Distribution function of $\log Y = \log X^2 = \log\left(\prod_{i=1}^{n} g_i^2\right)$ with Y given by Eq. (4.106) for different numbers of factors n. The histogram is taken over 100,000 samples and the solid line depicts the probability density function of a normal random variable with mean μn and standard deviation $\sigma\sqrt{n}$. Here, μ and σ denotes the mean and standard deviation of $\log g_i^2$ as given by Eq. (4.109).

Recall from Section 4.3.3 that one of the conditions needed for successful recovery via ALS is a lower bound on $\lambda_{\min}(B_j(v))$ with $B_j(v)$ from Eq. (4.102). In words, we need a lower tail bound on the smallest eigenvalue of sums of independent random operators of the form $Y \times |a\rangle\langle a|$, where $a \sim \mathcal{N}(0, \mathbb{1}_d)$ and Y is a product of squares of standard Gaussians, i.e.

$$Y = \prod_{i=1}^{n} g_i^2, \quad \text{with } g_i \sim \mathcal{N}(0,1). \tag{4.106}$$

We encountered a similar problem in Chapter 3, where the proof of the main Theorem 3.5 relied on the moment conditions in Definition 3.1. However, these standard concentration inequalities are not applicable to the problem of tensor recovery due to the large fluctuations of Y in Eq. (4.106): Because of the independence of the g_i, the k-th moments of Y are given by

$$\mathbb{E}Y^k = \prod_{i=1}^{n} \mathbb{E}g_i^{2k} = ((2k-1)!!)^n \,. \tag{4.107}$$

So although the expectation value of Y is 1 for all values of n, the variance of X scales as $(3!!)^n = 3^n$. This might be surprising since $Y > 0$ almost surely. The reason for the exponentially strong fluctuations about the mean are realizations of Y which are exponentially large, but which only occur with extremely small probability.

In order to illustrate the strong fluctuations of Y more clearly, note that

$$\log Y = \sum_{i=1}^{n} \log g_i^2. \tag{4.108}$$

That is, the logarithm of Y is a sum of independent random variables with

$$\mathbb{E}\log g_i^2 = \mu = -(\gamma + \log 2) \quad \text{and} \quad \mathbb{V}\log g_i^2 = \sigma^2 = \frac{\pi^2}{2}. \tag{4.109}$$

By the central limit theorem, we can approximate $\log Y$ for large n by a normal distribution with mean μn and standard deviation $\sigma\sqrt{n}$ as depicted in Fig. 4.6. In other words, Y is approximately log-normal distributed, and hence, the fluctuations of Y are of order $\sigma\sqrt{n}$ on a logarithmic scale, i.e. they manifest on the order of magnitude of Y. In contrast, the sum of n independent standard Gaussian is a Gaussian with standard deviation $\sqrt{n}$, which is naturally measured on a linear scale.

From the discussion above it becomes clear that standard tail bounds for sums of independent random variables are not suitable to bound $\lambda_{\min}(B_j(v))$ efficiently. For this purpose, a more thorough understanding of the distribution of the product of

squares of n independent Gaussian random variables is necessary. This was the main motivation behind the publications [SSK17b; SSK17a]. In this work, we derive a power-log series expansion of the cumulative density functions (cdf) of the following three random variables:

$$X = \prod_{i=1}^{n} g_i \qquad Y = \prod_{i=1}^{n} g_i^2 \qquad Z = \prod_{i=1}^{n} |g_i|, \tag{4.110}$$

based on the theory of special functions. More precisely, we show that the cdfs of said random variables can be expressed in terms of *Meijer G-functions*. Meijer G-functions are a family of special functions in one variable x that is closed under several operations including $x \mapsto -x$, $x \mapsto 1/x$, multiplication by x^p, differentiation, and integration. For the sake of completeness, we provide the definition of Meijer G-functions and further references in Appendix A.3. We refer the reader to [SSK17a] for the proofs of the following statements.

Lemma 4.8. *Denote by X, Y, and Z the product of N independent Gaussians, Gaussians squared, and their absolute value as defined in Eq.* (4.110)*. Define the function g_α by*

$$f_\alpha(z) := 1 - \frac{1}{2^\alpha} \cdot \frac{1}{\pi^{\frac{n}{2}}} G^{0,n+1}_{n+1,1}\left(z \middle| \begin{matrix} 1, 1/2, \dots, 1/2 \\ 0 \end{matrix} \right). \tag{4.111}$$

Here, $G^{0,n+1}_{n+1,1}$ denotes a Meijer-G function. Then, for any $t > 0$,

$$\mathbb{P}(X \leq t) = \mathbb{P}(X \geq -t) = f_1\left(\frac{2^n}{t^2}\right) \tag{4.112}$$

$$\mathbb{P}(Y \leq t) = f_0\left(\frac{2^n}{t}\right) \tag{4.113}$$

$$\mathbb{P}(Z \leq t) = f_0\left(\frac{2^n}{t^2}\right). \tag{4.114}$$

Based on the identities in Lemma 4.8, we now develop a power-log series of the cdfs of X, Y, and Z based on the theory of Fox H-functions.

Theorem 4.9. *Denote by X, Y, and Z the product of n independent Gaussians, Gaussians squared, and their absolute value, respectively, as given by Eq.* (4.110)*. Define the function*

$$f_{\nu,\xi}(u) := \nu + \frac{1}{2^\xi} \cdot \frac{1}{\pi^{n/2}} \sum_{k=0}^{\infty} u^{-1/2-k} \sum_{j=0}^{n-1} H_{kj} \cdot [\log u]^j \tag{4.115}$$

with

$$H_{kj} := \frac{(-1)^{nk}}{j!} \sum_{q=j}^{n-1} \left(\frac{1}{2}+k\right)^{-(q-j+1)} \times \sum_{j_1+\ldots+j_n=n-1-q} \prod_{t=1}^{n} \left\{ \sum_{\ell_1+\ldots+\ell_{k+1}=j_t} \frac{\Gamma^{(\ell_{k+1})}(1)}{\ell_{k+1}!} \left\{ \prod_{i=1}^{k-1} (k-i+1)^{-(\ell_i+1)} \right\} \right\} \tag{4.116}$$

Here, $j_i \in \mathbb{N}_0$ *and* $\ell_i \in \mathbb{N}$. *Then, for any* $t > 0$,

$$\mathbb{P}(X \leq t) = \mathbb{P}(X \geq -t) = f_{1/2,1}\left(\frac{2^n}{t^2}\right), \tag{4.117}$$

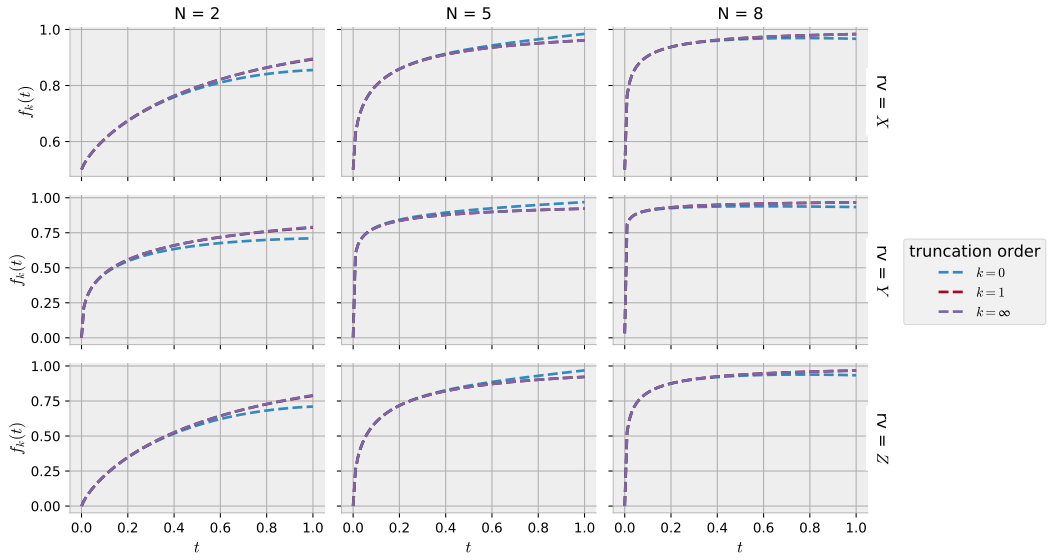
$$\mathbb{P}(Y \leq t) = f_{0,0}\left(\frac{2^n}{t}\right), \tag{4.118}$$

$$\mathbb{P}(Z \leq t) = f_{0,0}\left(\frac{2^n}{t^2}\right). \tag{4.119}$$

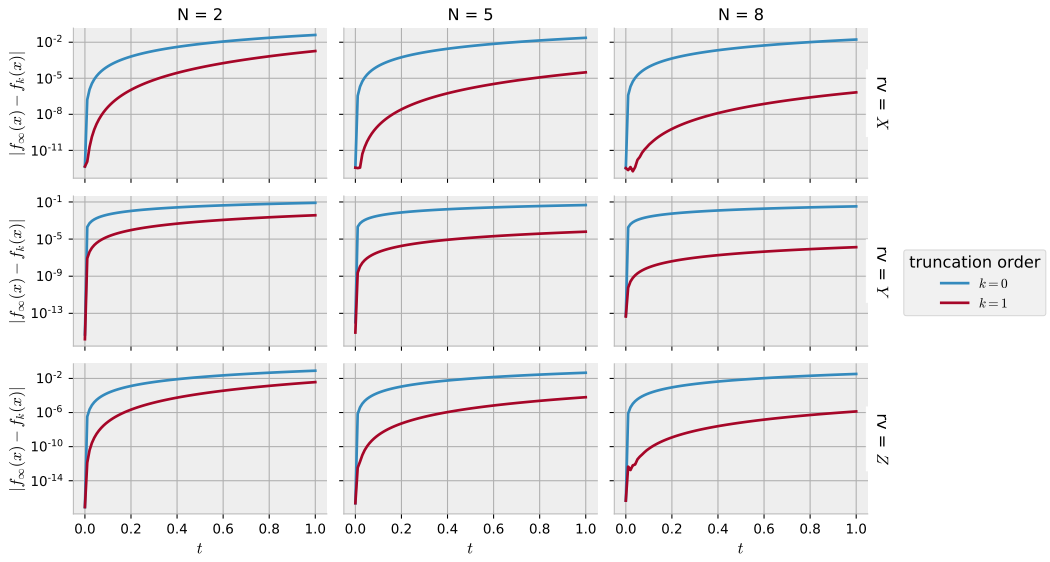
The proof can be found in [SSK17a]. Here, we are mostly interested in the power-log series expansion (4.115), because already its zeroth and first order truncation approximate the cdfs of X, Y, and Z well as shown in Fig. 4.7. Said figure depicts the true cdf as well as their series approximations from Eq. (4.115) for $k = 0$ and $k = 1$. We can see that for all three random variables as well as for all values of n, the zeroth order approximation is good for small values of t, but worsens for increasing t. The first order approximation behaves similarly, but the approximation error remains small enough so that it cannot be distinguished from the true cdf in the linear plots. Therefore, we show the logarithmic approximation error in Fig. 4.7b, where we see that the first-order approximation error remains below 10^{-2} for all shown cases. Note that the approximation also gets better for increasing n. Therefore, we obtain the worst approximations for the generally better behaved smaller values of n.

In conclusion, Fig. 4.7 shows that the power-log series expansion of the cdfs of the random variables X, Y, and Z provides excellent approximation the very truncation order $k = 1$. The problem of proving such a statement rigorously remains, however. Furthermore, applying this approximation to prove the concentration properties from Section 4.3.3 is also left for future work.

For the remainder of this section, we investigate the distribution and scaling of the crucial constant $\frac{\delta_B}{\delta_G}$ introduced in Section 4.3.3 w.r.t. the parameters N and d . Recall the main result of Section 4.3.3, the condition (4.99), which relates the quotient of the measurement constants $\frac{\delta_B}{\delta_G}$ to the error δ_I in the initialization. We are now going to investigate the scaling of said fraction numerically for the Gaussian measurement



(a) CDFs and approximations



(b) Approximation errors

Figure 4.7.: Figure 4.7a shows the CDFs ($k = \infty$) of the random variables X, Y, and Z with their power-log series (4.117)–(4.119), respectively, truncated at different orders k. Note that the first order approximation $k = 1$ lies right on top of the true value, and hence, is not visible in this plot. In Fig. 4.7b, we show the approximation error, i.e. the absolute value of the difference of the truncation and the true value, on a log-scale.

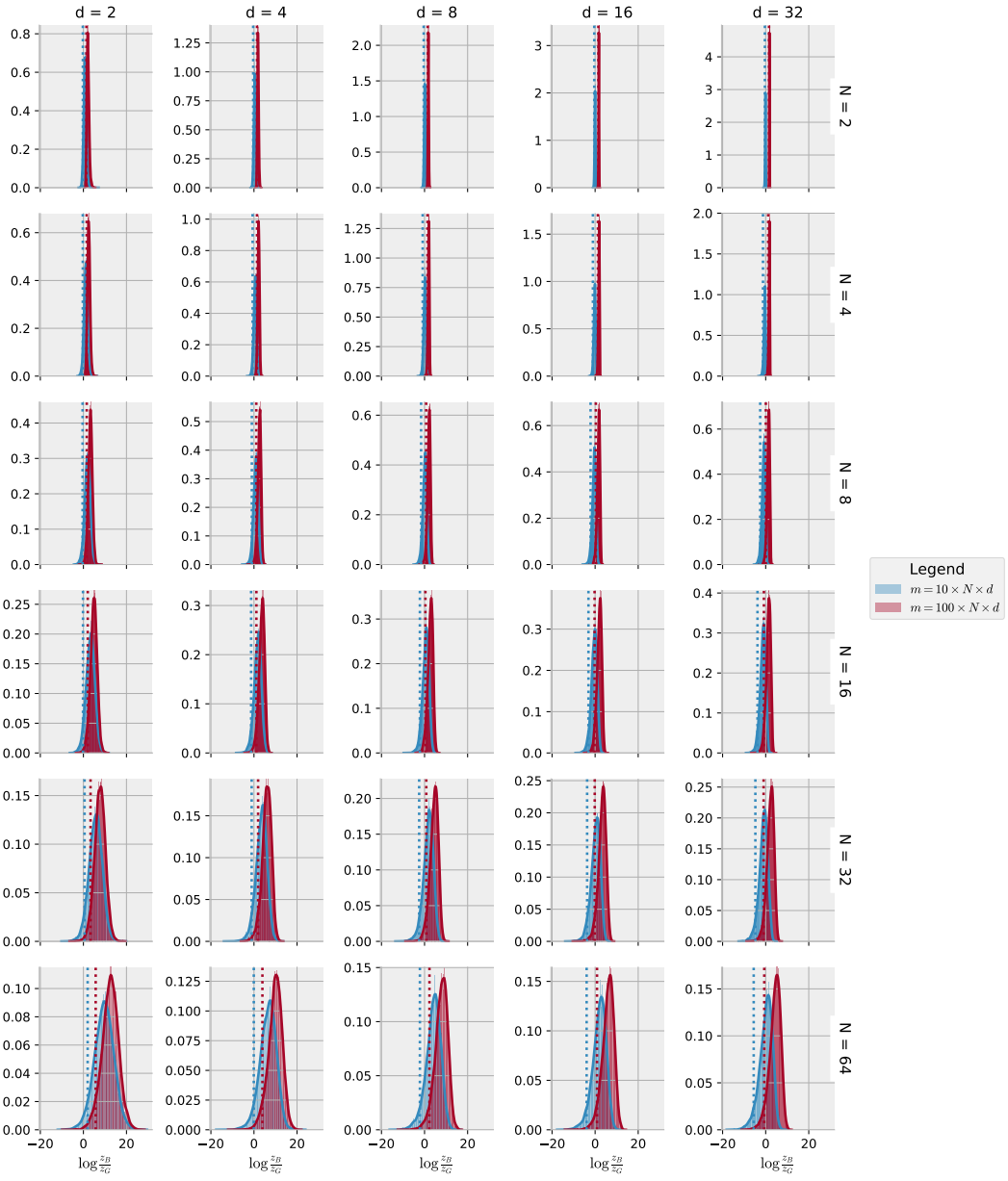


Figure 4.8.: Empirical distribution of $\frac{Z_\mathrm{B}}{Z_\mathrm{G}}$ for Gaussian product measurements. Here, Z_B and Z_G are given by Eq. (4.120). The solid lines indicate the smoothed histograms over 10000 samples and the dotted lines their 0.05 quantiles, i.e. 5 % of the samples are smaller than the value shown. For each combination of the number of sites N and local dimension d, we choose the number of measurements $m = CNd$. For the blue curve we have $C = 10$ and for the red curve $C = 100$.

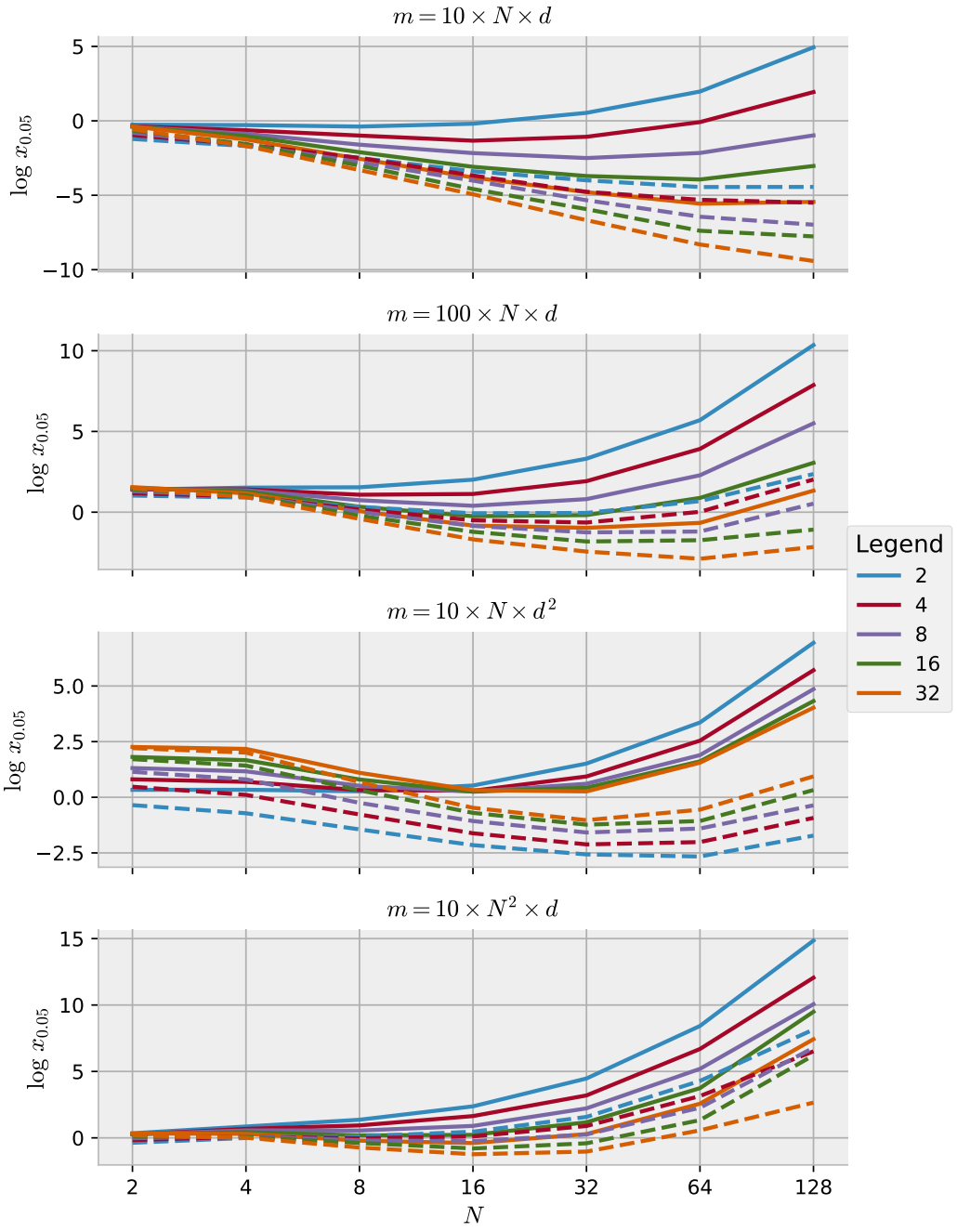


Figure 4.9.: Quantiles of the random variable $\frac{Z_B}{Z_G}$ for Gaussian measurements. The solid lines indicate the empirical 5 % quantile $x_{0.05}$. The dotted lines show the fractions of the "uncorrelated" quantiles of Z_B and Z_G, i.e. $\frac{x_{B,0.025}}{x_{G,0.975}}$. Therefore, the dotted lines represent a lower bound on $x_{0.05}$, which might be easier to proof.

ensemble. For this purpose, we sample the operators $B_j(v)$ and $G_j(v, v^\perp)$ defined in Eqs. (4.49) and (4.51), respectively for Gaussian measurements and for different values of N, d, and m. In the following, we use the shorthand notation

$$Z_{\mathrm{B}} = \lambda_{\min}\left(B_j(v)\right), \quad \text{and} \quad Z_{\mathrm{G}} = \left\| G_j(v, v^\perp) \right\|_{2\to 2}. \tag{4.120}$$

In Fig. 4.8, we show the empirical distribution of $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ over 10000 realizations. Here, we chose the number of measurements according to $m = CNd$ for $C \in \{10, 100\}$, which is the same asymptotical scaling as the information-theoretic lower bound Nd. The most influential parameter on the empirical distribution of $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ is the number of sites N. From Fig. 4.8, we see with that increasing N, the distribution becomes more spread out. With Fig. 4.6 in mind, this is to be expected as the strong fluctuations of Z_{B} and Z_{G} are due to their weighting factors, which are products of independent Gaussian. The dependence of the empirical distribution of $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ on d is weak for the depicted parameter range, as the histograms look very similar in each row.

The dotted lines in Fig. 4.8 shows the 0.05 quantile $x_{0.05}$ of each empirical distribution. In general, for a random variable X the q-th quantile is defined as the smallest number x_q such that

$$\mathbb{P}\left(X \leq x_q\right) \leq q. \tag{4.121}$$

Therefore, the conditions on the measurements (4.49) and (4.51) hold with probability 0.95 for $\frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}} = \tilde{x}_{0.05}$, if we denote by $\tilde{x}_{0.05}$ the quantile of the true distribution. Here, we approximate the latter by the quantile of the empirical distribution over 10000 samples.

To further investigate the scaling of the quantiles, we plot them as a function of N in Fig. 4.9 for different values of d and for different sampling rates m. In general, for the parameters under consideration, the quantiles either grow monotonically, or do so after attaining their minimum value for some value of N. In other words, the plots indicate that for fixed d and $m > 10Nd$, the quantiles are bounded from below. Furthermore, comparing the plots with linear scaling for $C = 10$ and $C = 100$ shows that, as expected, increasing the constants also increases the quantiles throughout. We also get larger quantiles for the other two quadratic sampling rates.

Finally, the dashed lines in Fig. 4.9 show the lower bound from combining separate quantiles for Z_{B} and Z_{g} incoherently: Note that if $Z_{\mathrm{B}} \geq \delta_{\mathrm{B}}$ with probability $1 - \frac{q}{2}$ and $Z_{\mathrm{G}} \leq \delta_{\mathrm{G}}$ with probability $1 - \frac{q}{2}$, then a union bound argument shows that

$$\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}} \geq \frac{\delta_{\mathrm{B}}}{\delta_{\mathrm{G}}} \tag{4.122}$$

with probability $1 - \frac{q}{2} - \frac{q}{2} = 1 - q$. Therefore, we can lower bound the q-th quantile of $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ by the fraction of the $\frac{q}{2}$ and $1 - \frac{q}{2}$ quantiles of Z_{B} and Z_{G}, respectively. This

lower bound indicated by the dashed lines is exactly used by the proof strategy with separate probabilistic bound of Z_{B} and Z_{G}.

Note that the dotted lines show a similar behavior as the solid lines, and in particular, they are lower bounded for all values depicted as well. Hence, these numerics suggest that such an approximation of the quantiles of $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ is might be able to provide suitable analytic bounds. In conclusion, Fig. 4.9 indicates that the crucial random variable $\frac{Z_{\mathrm{B}}}{Z_{\mathrm{G}}}$ can – at least for fixed local dimension d – be lower bounded by a constant.

4.3.5. Numerical reconstruction

In this section, we are going to numerically demonstrate the recovery of low-rank – and not unnecessarily rank one – tensors via ALS. For this purpose, we introduce a modified version of the ALS algorithm and also comment on implementation details, which enable the reconstruction of large tensors.

The ALS algorithm as introduced in Alg. 1 has two drawbacks in practice: On the one hand, it uses a fresh batch of size m of measurements for each micro-iteration. Therefore, it requires HNm measurements and measurement tensors. On the other hand, it does not provide an explicit initialization X_{init}. These choices were made to simplify the analysis of the algorithm, but they make Alg. 1 less suited in practice.

For the numerical experiments in this section, we are going to use a modified ALS algorithm. To alleviate the drawbacks mentioned above, we are simply going to reuse the same batch of size m of measurements in each iteration. Furthermore, we use these measurements to compute a suitable initialzation: Consider the problem of recovering a rank r tensor $X \in \mathbb{R}^{d^N}$. Let

$$\tilde{X}_{\mathrm{init}} = \sum_{l=1}^{m} b^{(l)} A^{(l)}, \tag{4.123}$$

then we use a rank r approximation of $\tilde{X}_{\mathrm{init}}$ as initialization, that is

$$X_{\mathrm{init}} = \texttt{compress}_r(\tilde{X}_{\mathrm{init}}). \tag{4.124}$$

Here, $\texttt{compress}_r$ denotes a suitable compression routine such as SVD-compression introduced in Section 4.1.2 yielding a rank r approximation. To understand the motivation behind Eq. (4.123), note that $\tilde{X}_{\mathrm{init}} = \frac{1}{m}\mathcal{A}^\dagger \mathcal{A} X$, where $\mathcal{A}$ is the measurement operator introduced in Proposition 3.3

$$\mathcal{A} = \sum_{l=1}^{m} |e_l\rangle\langle A^{(l)}| \tag{4.125}$$

with e_l denoting the l-th canonical basis vector. Therefore, for the Gaussian measurements introduced in Section 4.3.4, we have

$$\begin{aligned}
\mathbb{E}\mathcal{A}^\dagger\mathcal{A} &= \mathbb{E}\left(\sum_l \left|A^{(l)}\right\rangle\left\langle A^{(l)}\right|\right) && (4.126)\\
&= m\mathbb{E}\left(\otimes_{i=1}^N |a_i\rangle\langle a_i|\right) && (4.127)\\
&= m\otimes_{i=1}^N \left(\mathbb{E}\,|a_i\rangle\langle a_i|\right) && (4.128)\\
&= m\mathbb{1}_d \otimes \cdots \mathbb{1}_d, && (4.129)
\end{aligned}$$

and hence $\mathbb{E}\tilde{X}_{\text{init}} = X$. In words, in expectation, the (uncompressed) initialization is exactly the tensor we are trying to reconstruct. Therefore, we expect that $\tilde{X}_{\text{init}}$ is close to X provided m is "large enough". The exact value of m to make this precise depends on the fluctuations of $\mathcal{A}$.

Additionally, we also employ a left-right-left sweep as suggested by the analysis in Section 4.3.3. The "practical" ALS algorithm used throughout this section is

summarized below:

Algorithm 2: Practical Alternating Least Squares (ALS) for ℓ_2 minimization

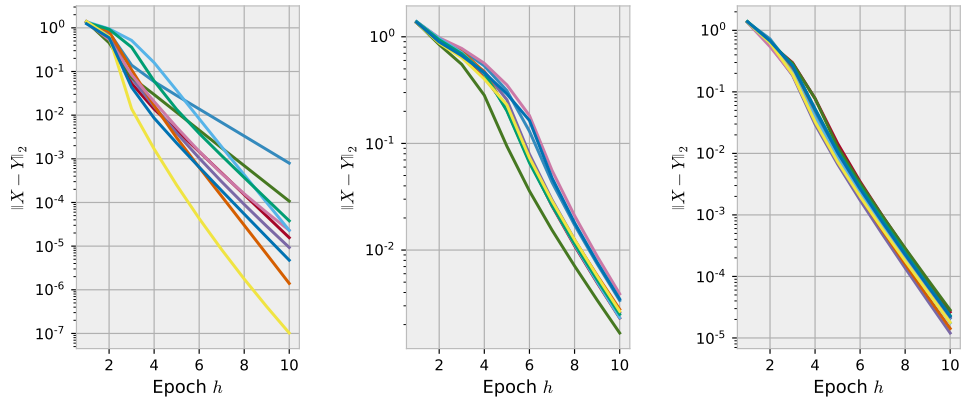
Input : Number of epochs H, target rank r, measurement tensors $A^{(l)}$ and measurement outcomes $b^{(l)}$ with $i = 1, m$

```
/* Spectral initialization                                          */
```
$Y \leftarrow \texttt{compress}_r(\sum_{l=1}^{m} b^{(l)} A^{(l)})$ `right_canonicalize`(Y)
for $h \leftarrow 1$ **to** H **do**
  for $n \leftarrow 1$ **to** N **do**
    for $l \leftarrow 1$ **to** m **do**
      `/* contract` $A^{(l)}$ `with all but n-th local tensors */`
      $B^{(l)} \leftarrow \texttt{contract}(A^{(l)}, Y_{[N]\setminus n})$
    $\hat{Z} \leftarrow \text{argmin}_Z \sum_l \left(b^{(l)} - B^{(l)} Z\right)^2$
    `/* update the n-th local tensor inplace */`
    $Y_n \leftarrow \texttt{left\_normalize}(\hat{Z})$
  for $n \leftarrow N$ **to** 1 **do**
    for $l \leftarrow 1$ **to** m **do**
      `/* contract` $A^{(l)}$ `with all but n-th local tensors */`
      $B^{(l)} \leftarrow \texttt{contract}(A^{(l)}, Y_{[N]\setminus n})$
    $\hat{Z} \leftarrow \text{argmin}_Z \sum_l \left(b^{(l)} - B^{(l)} Z\right)^2$
    `/* update the n-th local tensor inplace */`
    $Y_n \leftarrow \texttt{right\_normalize}(\hat{Z})$

Output: Y

In Fig. 4.10, we show the reconstruction error of Alg. 2 as a function of the epoch h for different ranks of the target X as well as different sampling rates m. The target tensor was chosen randomly and is kept fixed in each figure. Each line corresponds to a different set of randomly chosen measurement tensors $A^{(l)}$ from the Gaussian ensemble.

Figure 4.10a depicts the reconstruction error for the recovery of a rank 1 tensor X from $m = 0.05d^N$ measurements. The convergence of the estimate to the true value is very fast as about six sweeps are enough for most runs to reduce the error below 10^{-2}. Since the reconstruction error decays approximately linear on a semi-logarithmic scale, the convergence in these examples is exponentially fast. We see a slower convergence of the reconstruction error for a rank-10 tensor in Fig. 4.10b. This is expected as the number of parameters for a rank-r MPS scale as $\mathcal{O}(Ndr^2)$, and hence, more measurements are necessary to recover such a tensor successfully. By doubling mm in Fig. 4.10c, we are able to speed up convergence and obtain a



(a) $r = 1$, $m = 0.05\, d^N$ (b) $r = 10$, $m = 0.1\, d^N$ (c) $r = 10$, $m = 0.2\, d^N$

Figure 4.10.: Recovery error of ALS (Algorithm 2) as a function of the epoch h. All pictures show the recovery of a fixed randomly chosen tensor $X \in \mathbb{R}^{d^N}$ for $d = 4$ and $N = 8$. Each line corresponds to a different set of Gaussian product measurements. In Fig. 4.10a, we chose X to be of rank 1, whereas Figs. 4.10b and 4.10c show the reconstruction of a rank 10 tensor. For initialization, we used the spectral initialization (4.124).

similar speed as for the rank-1 case. Note that the reconstruction error curves for the rank-10 tensor are more uniform for different measurement tensors. This is likely due to correlations between local tensors for non-product tensors.

In all the examples shown, we are able to recover the target tensor with only few iterations from fewer measurements than d^N. By exploiting the low-rank structure, we are not only able to reduce the sample complexity, but also speed up the computation.

For the remainder of this section, we comment on implementation details, which enable scaling Alg. 2 even to large instances. The first bottleneck in a straightforward implementation of the ALS algorithm is the initialization in line 1 of Alg. 2. A naïve implementation in MPNUM would read

```
A = [mp.random_mpa(N, d, rank=1) for _ in range(m)]
b = [mp.inner(a, X) for a in A]
X_init, _ = mp.sumup(A, weights=b).compression(rank=r)
```

The first two lines are simply generating the random measurement tensors `A` and the corresponding measurement values `b`. Here, `X` is the original tensor of rank r, we are trying to recover. The third line first computes $\tilde{X}_{\text{init}}$ from Eq. (4.123) and then compresses it to the desired rank. Note that as a sum of m rank-1 operators,

$\tilde{X}_{\text{init}}$ is of rank m. The main bottleneck here is bringing $\tilde{X}_{\text{init}}$ to canonical form as required for the compression: A general-purpose canonicalization as implemented in MPNUM has to compute matrix products of $md \times m$ and $m \times dm$ matrices in this case. Furthermore, the subsequent compression is based on SVDs of similar sized matrices.

In order to speed this up, we implemented a joint sum-and-compression routine in MPNUM, which benefits from the explicit representation of sums of rank-1 tensors. The corresponding local tensors are highly sparse, and up to a constant factor, already normalized. The last line of the example above then translates to

```
X_init = mp.special.sumup(A, rank=r, weights=b)
```

which provides a 5× speedup for the computations in Fig. 4.10c and reduces memory footprint significantly.

For the iteration steps of Alg. 2, the main bottleneck for large instances is the computation of the matrices B in the lines 5 and 10. These can be speed up in two different ways: First, not all contractions in Eq. (4.38) have to be recomputed in every step. In a left-sweep, the contraction at the i-th site are reused i times until said site is update. Furthermore, once the site is updated, its contraction is required for the following $N - i - 1$ updates. Therefore, by using dynamic programming techniques and caching intermediate computation, which are reused later, we are able to speed up the computation of the B significantly.

Second, we note that the computation of each row $B^{(l)}$ of B is independent from all the other rows, and therefore, can be parallelized on shared-memory architectures. In particular, since this is a typical example of a same-instruction-multiple-data parallelism (SIMD), it is suitable for implementation on modern GPUs. The latter are especially suited for problems with number of independent SIMD computations much larger than the number of GPU cores, and which do not require a lot of data copying between the CPU and the GPU. Both points apply in our case: For the rather small example of Fig. 4.10c, we already have $m = 13107$, which is significantly larger than 3584 – the number of cores of the state-of-art NVIDIA Pascal P100 GPU accelerator. Since each micro-iteration only requires updating the previous local tensor of size $r^2 d$ on the GPU and downloading the matrix B of size $m \times r^2 d$ to the CPU, the faster computation of B on the GPU amortizes already for medium-sized instances and provides speedups on the order of 10× even on older GPUs.

We provide optimized implementations of the ALS algorithm 2 for both CPU and GPU at `https://github.com/dseuss/pycsalgs`.

4.4. Conclusion & outlook

The MPS tensor format provides an efficient representation for many tensors occurring in practical applications. In contrast to general tensors, which require exponen-

tially many parameters w.r.t. the order of the tensor, tensors of MPS rank r can be parameterized using only $\mathcal{O}(Ndr^2)$ real numbers. Here, we show how to recover such an efficient representation from few linear measurements using the ALS algorithm. In contrast to previous work, which either requires an exponentially large number of measurements or incompressible Gaussian measurement tensors, our work investigates recovery using product measurements. Therefore, the proposed scheme aims no only to be sample efficient, but also computationally efficient.

The main result of this chapter is the sufficient condition for reconstruction from rank-one measurements via ALS in Eq. (4.99). It relates the fraction $\frac{\delta_\mathrm{B}}{\delta_\mathrm{G}}$ of constants quantifying concentration properties of the measurements and the deviation of the initialization from the true value. As discussed below said equation, there is a trade-off between the properties of the measurements and the initialization: A tighter bound for the conditions (4.50) and (4.52) allows us to weaken the requirements on the initialization, and vice versa. In Section 4.3.4, we examine the conditions for Gaussian product measurements. These results suggest that for fixed d, $m = \mathcal{O}(N)$ measurements are indeed sufficient to fulfill the conditions with high probability. Furthermore, we also demonstrate successful numerical reconstruction in Section 4.3.5 of rank-one and low-rank tensors with severe under-sampling. To support the numerical experiments, we developed MPNUM – a library for MPS representation algorithms.

The most pressing goal for future work is to proof a lower bound for the fraction $\frac{\delta_\mathrm{B}}{\delta_\mathrm{G}}$ for Gaussian rank-one measurements. For this purpose, we have computed a power-log series expansion for the cdfs of products of independent Gaussians and their square. Numerical experiments have shown that the low order truncations of this series provide extremely accurate approximations. Therefore, one viable option to bound the measurement constants is to use the zeroth- or first order truncation of the power-log series expansion and bound the error of the truncation. An alternative approach is based on the observation that sums of products of independent Gaussians are dominated by a single, namely the largest summand. This is due to the fact that products of Gaussians exhibit fluctuations on orders of magnitude instead of on a linear scale. Making this precise would allow for a tremendous simplification of the problem as the full sum can be replaced by a sum over few terms.

Another important point for future work is the generalization of the convergence proof to higher-rank target tensors X. As mentioned in Section 4.3.3, the setting of higher-rank matrices is analysed by means of the principle angle distance in [ZJD15]. This, or a similar distance measure for the local tensors, should also be useful for the higher-rank tensor case.

We also note that the proof in of Theorem 4.7 is certainly not optimal. Strengthening the condition for the initialization from a local condition, i.e. a condition for each local tensor, to a global condition, i.e. a condition for the full initialization tensor, would make the result more robust. However, note that the ALS algorithm does not

necessarily improve the error of the full tensor in each micro-iteration: Consider the case, where the first local tensor equal to its true value, i.e. $y_1^0 = x_1$, while the other local tensor deviate from their true values. Then, the exact form of the minimizer in Lemma 4.5 shows that the update y_1^1 is not equal to x_1 any longer. This increases the global error $\|Y - X\|_2$ in the first micro-iteration. One way to bound the possible error in each micro iteration is to replace the exact minimization of the empirical ℓ_2-error in each step by a finite gradient descent step.

Finally, a better understanding of the initial value is an important point for future work. Note that the initialization used in Section 4.3.5 is based on the following naïve intuition: First, the uncompressed initialization $\tilde{X}_{\text{init}}$ defined in Eq. (4.123) should be "close" to the true value[11]. Second, the compression of $\tilde{X}_{\text{init}}$ to the appropriate rank should not increase the error too dramatically. One way to bound the error caused by compression is Eq. (4.22). However, in numerical experiments, we note that after re-normalization, X_{init} is often closer to X than $\tilde{X}_{\text{init}}$, especially for very large N. This is not surprising as the compression regularizes the initial estimate – this is exactly the effect that allows for recovering X from fewer than d^N measurements. A better understanding of these effects and possibly more elaborate initialization schemes would further reduce the number of required measurements for reconstruction.

[11] Note that close is very vague in this context: In principle, the local condition in Theorem 4.7 still allows for an exponentially small overlap of X and X_{init}.

5. Conclusion

To summarize, we have investigated three inference problem from quantum physics, which are subject to different types of constraints. The main motivation of this work stems from the observation that exploiting additional structure in inference problems often helps to reduce their sample complexity. However, the examples in this work differ in how taking the constraints into account affects their computational complexity.

The first inference problem under consideration is quantum state estimation – reconstructing the density matrix of a quantum system from measurements. More specifically, we are interested in optimal error regions that take the positive semi-definite constraint of physical states into account. For this purpose, we show that deciding whether an ellipsoid is contained in the set of positive semi-definite matrices is **NP**-hard. As a consequence, computing the radius of optimal Bayesian credible ellipsoids for QSE is computationally intractable, whereas the unconstrained problem can be solved efficiently. For Frequentist confidence regions, this result implies that computing any property of truncated confidence ellipsoids that is sensitive to truncation is hard as well. In conclusion, although there are settings where taking into account the physical constraints of QSE drastically improves the power of the error region, doing so in an optimal way is computationally intractable.

Note that this work does not preclude the existence of algorithms for uncertainty quantification in QSE that work well-enough in practice. Our hardness results relies on strong assumptions, some of which might be relaxed for practical applications. For example, our results leave room for the existence of efficient approximate solutions. Rather, our hardness result should be understood as an absolute upper bound on what such algorithms can achieve.

In Chapter 3, we investigate characterizing linear optical circuits and the related phase retrieval problem. To overcome to challenge of phase-insensitive measurements, we map the problem to rank-one matrix recovery. By exploiting the exact rank-one constraint, we are able to perform reconstruction using an asymptotically optimal number of measurements. Furthermore, our recovery protocol can be implemented efficiently using a positive-semidefinite program called "PhaseLift" and it is robust to noise as the rigorously proven recovery guarantees show.

We also propose a measurement ensemble for phase retrieval tailored to the application in optics. In contrast to the Gaussian ensemble used in previous work,

the RECR ensemble only necessitates the ability to prepare four complex phases per mode and discrete magnitudes. This allows for calibrating the preparation stage more accurately, and hence, reduce the total error due to a mismatch of theoretical and implemented input vectors. Using a unified proof strategy, which was developed for this work, as well as numerical experiments we are able to show that the RECR ensemble's performances matches the well-established Gaussian scheme.

From an experimentalist's point of view, characterization of linear-optical networks via PhaseLift is favourable because it reduces the number of different measurement configurations required. As reconfiguring the currently used hardware to prepare another input takes more time than the actual measuring process, the sample efficiency of our protocol reduces the total amount of time required.

The problem of low-rank tensor reconstruction shows that in some cases constraints are necessary for an efficient solution. High-order tensors are hard to deal with computationally due to the exponential scaling of the number of parameters. This motivated the development of different tensor formats such as the MPS format considered here, which by construction reflects the correlation structure of certain tensors occurring in applications. Therefore, they allow for an efficient representation of these relevant tensors. Here, we answer the question whether such a tensor with efficient MPS representation can be reconstructed from few linear measurements. In contrast to previous work, we are interested in both sample efficiency and computational complexity. Hence, we consider product measurements, which are efficiently representable as well.

The analysis of the ALS algorithm yields a sufficient condition that guarantees successful recovery of any rank-one tensor using rank-one measurements. As a prototypical example, we consider Gaussian product measurements, which are numerically shown to satisfy these conditions for a large variety of parameters. Additionally, numerical reconstruction experiments show that we are able to reconstruct large tensors from serenely under sampled measurements.

The ALS algorithm using rank-one measurements combines both sample and computational efficiency for tensor reconstruction. By exploiting the low-rank constraint we obtain an exponential improvement for the sampling rate compared to naïve approaches under the assumptions stated. This reduction of the number of measurements necessary for recovery is also crucial for making the reconstruction computationally efficient. The reduction of sample and computational complexity for low-rank tensor recovery is For the latter, reconstruction without additional constraints is still feasible and, at least in the sense of polynomial scaling, still efficient. Future work is necessary to prove the recovery condition for certain measurement ensembles. The most promising – as the numerical investigation in this chapter shows – are Gaussian product measurements.

What these three estimation problems have in common is the fact that the additional structure present in the form of constraints can be used to improve the quality of the estimate in principle. However, the ability to make use of this in practice highly depends on the problem. On the one hand, exploiting those constraints optimally constitutes a computationally hard problem in the case of uncertainty quantification for quantum state estimation. On the other hand, the problem of efficiently reconstructing tensors only becomes feasible due to this additional structure as the embedding space grows exponentially as a function of the order of the tensor. Phase retrieval and, more generally, low-rank matrix recovery lies somewhere between those two extreme cases as the sample complexity is reduced only by a linear factor at the cost of a slightly higher – but still efficient – computational complexity.

A. Appendix

A.1. Generalized Bloch representation

Here, we provide the particular generalizations σ_i of the Pauli matrices used in Sec. 2.4.2. These are exactly the generators of the group $\mathrm{SU}(d)$, see e.g. [Kim03; BK03] for more details. Denote by $\{|i\rangle\}_i$ an orthonormal basis and let

$$\begin{aligned}
\Xi_{jk}^{(\mathrm{Re})} &= |j\rangle\langle k| + |k\rangle\langle j|, \\
\Xi_{jk}^{(\mathrm{Im})} &= -\mathrm{i}\Big(|j\rangle\langle k| - |k\rangle\langle j|\Big), \\
\Xi_{l}^{(\mathrm{diag})} &= \sqrt{\frac{2}{l\,(l+1)}}\left(\sum_{j=1}^{l} |j\rangle\langle j| - l|l+1\rangle\langle l+1|\right).
\end{aligned}$$

We now define the generalized Pauli matrices in terms of these auxiliary matrices:

$$\{\sigma_i : i = 1, \ldots, i_d\} = \left\{\Xi_{jk}^{(\mathrm{Re})} : 1 \leq j < k \leq d\right\}, \tag{A.1}$$

$$\{\sigma_i : i = i_d + 1, \ldots, 2i_d\} = \left\{\Xi_{jk}^{(\mathrm{Im})} : 1 \leq j < k \leq d\right\}, \tag{A.2}$$

$$\left\{\sigma_i : i = 2i_d + 1, \ldots, d^2 - 1\right\} = \left\{\Xi_{l}^{(\mathrm{diag})} : 1 \leq l \leq d - 1\right\}, \tag{A.3}$$

where $i_d = d(d-1)/2$. Note that the elements of the sets in Eq. (A.1), (A.2), and (A.3) generalize the Pauli matrices σ_{X}, σ_{Y}, and σ_{Z}, respectively. Since only this structure is crucial to our proof, the order of the elements in Eq. (A.1)–(A.3) in arbitrary, and hence, the definition in terms of sets is well defined for our purposes.

A.2. Experimental details

A.2.1. Reference Reconstructions

Since our goal is to benchmark the PhaseLift characterisation technique, and not the performance of the optical chip, we compare the experimental reconstructions to reconstructions obtained with a different technique. These reference reconstructions are obtained in two steps. First, we estimate the absolute value of each component from single photon data: From Eq. (3.4), we see that by inserting single photons into

Dimension n	2	3	5
Gaussian	20	30	40
RECR	6	31	39

Table A.1.: Total number of preparation vectors taken during experiment.

the k-th input of the device – that is choosing the standard basis vectors as inputs $\alpha = e_k$ – we can estimate $|M_{i,k}|$. For each input port, the counts of each detector are normalised to take into account the detector efficiencies and then divided by the total of the counts in all detectors. The square roots of these numbers are used as the estimated amplitudes of the matrix elements. Second, we estimate the phase of each component using HOM-dips [HOM87], following a similar approach to [LO12; Dha+16]. However, this second step is time-consuming and only reliable for small devices. Therefore, for the larger devices, we only perform the first step and compare only the magnitudes of the matrix elements of the reconstructions.

A.2.2. Data analysis

As mentioned in the main text, we estimate the intensity measurements from single photon counting rates. After correcting for detector efficiency, all counting rates are scaled by a constant such that the resulting intensities obey $\max_l \sum_j I_j(\alpha^{(l)}) = 1$. This only amounts to scaling the transfer matrix by a constant, which does not influence the end result since we later rescale the obtained reconstruction appropriately (see Eq. (A.4)). However, this simple rescaling helps with numerical stability in the SDP solver. We provide a ready for use implementation of the PhaseLift convex program (3.12) as well as related algorithms in the open source library PYPLLON [Sue17b],

The post-processing of a reconstruction $M^\sharp$ consists of two steps: First, we rescale the reconstruction by a constant such that

$$\max_i \|(M^\sharp)_i\|_{\ell_2} = 1, \tag{A.4}$$

where $(M^\sharp)_i$ denotes the i-th row of $M^\sharp$. In an ideal experiment, $M^\sharp$ would be unitary and, therefore, every row would have unit norm. However, due to loss in the characterised circuit as well as detector inefficiencies, the norm of each row is smaller than one. Since we cannot distinguish the two sources of loss in our current experimental setup, we cannot characterise the absolute photon loss in the circuit, but only the relative losses of the rows. Estimating the dark counts in future experiments would enable characterising the absolute photon loss in the circuit as well.

The second post-processing step consists of fixing phases of the reconstructions: Recall that we are only able to recover the transfer matrix up to its row phases since the global phases of the rows are lost in the intensity measurements. Therefore, we

fix the row phases of the PhaseLift reconstructions in Fig. 3.5 by minimizing the Frobenius distance to the target unitary and compute the error as

$$\min_{\mu:|\mu_i|=1} \left\| M_{\text{target}} - \text{diag}(\mu) M^\sharp) \right\|_2 \tag{A.5}$$

However, since the HOM-dip reconstruction is insensitive to global phases of the columns as well, we have to minimize both row and column phases for the HOM-dip reconstructions in Fig. 3.5. Furthermore, since in Fig. 3.4 the HOM-dip reconstruction is taken as reference value, we have to minimize the row and column phases for all PhaseLift reconstructions in that picture as well. The raw data as well as the analysis scripts are going to be made available at `https://github.com/dseuss/phaselift-paper`.

A.3. Meijer G-functions

Meijer G-functions are a family of special functions in one variable that is closed under several operations including

$$x \mapsto -x,\ x \mapsto 1/x,\ \text{multiplication by } x^p,\ \text{differentiation, and integration.} \tag{A.6}$$

Definition A.1. *For integers m, n, p, q satisfying $0 \le m \le q$, $0 \le n \le p$ and for $a_i, b_j \in \mathbb{C}$ (with $i = 1, \dots, p$; $j = 1, \dots, q$), the Meijer G-function $G^{m,n}_{p,q}\left(\cdot \middle| \begin{matrix} a_1, a_2, \dots a_p \\ b_1, b_2, \dots, b_q \end{matrix} \right)$ is defined by the line integral*

$$G^{m,n}_{p,q}\left(z \middle| \begin{matrix} a_1, a_2, \dots a_p \\ b_1, b_2, \dots, b_q \end{matrix} \right) = \frac{1}{2\pi} \int_{\mathcal{L}} \mathcal{H}^{m,n}_{p,q}(s) z^{-s}\ , \tag{A.7}$$

with

$$\mathcal{H}^{m,n}_{p,q}(s) := \frac{\prod_{j=1}^{m} \Gamma(b_j + s) \prod_{i=1}^{n} \Gamma(1 - a_i - s)}{\prod_{i=n+1}^{p} \Gamma(a_i + s) \prod_{j=m+1}^{q} \Gamma(1 - b_j - s)}\ . \tag{A.8}$$

Here,

$$z^{-s} = \exp\left(-s\{\log|z| + \arg z\}\right), \quad z \neq 0, \quad = \sqrt{-1}, \tag{A.9}$$

where $\log|z|$ represents the natural logarithm of $|z|$ and $\arg z$ is not necessarily the principal value. Empty products are identified with one. The parameter vectors a and b need to be chosen such that the poles

$$b_{j\ell} = -b_j - \ell \quad (j = 1, 2, \dots, m;\ \ell = 0, 1, 2, \dots) \tag{A.10}$$

of the gamma functions $s \mapsto \Gamma(b_j + s)$ and the poles

$$a_{ik} = 1 - a_i + k \quad (i = 1, 2, \dots, n;\ k = 0, 1, 2, \dots) \tag{A.11}$$

of the gamma functions $s \mapsto \Gamma(1 - a_i - s)$ *do not coincide, i.e.*

$$b_j + \ell \neq a_i - k - 1 \quad (i = 1, \ldots, n;\ j = 1, \ldots, m;\ k, \ell = 0, 1, 2, \ldots). \tag{A.12}$$

The integral is taken over an infinite contour $\mathcal{L}$ *that separates all poles* $b_{j\ell}$ *in Eq.* (A.10) *to the left and* a_{ik} *in Eq.* (A.11) *to the right of* $\mathcal{L}$*, and has one of the following forms:*

1. $\mathcal{L} = \mathcal{L}_{-\infty}$ *is a left loop situated in a horizontal strip starting at the point* $-\infty + \phi_1$ *and terminating at the point* $-\infty + \phi_2$ *with* $-\infty < \phi_1 < \phi_2 < +\infty$*;*
2. $\mathcal{L} = \mathcal{L}_{+\infty}$ *is a right loop situated in a horizontal strip starting at the point* $+\infty + \phi_1$ *and terminating at the point* $+\infty + \phi_2$ *with* $-\infty < \phi_1 < \phi_2 < +\infty$*;*
3. $\mathcal{L} = \mathcal{L}_{\gamma\infty}$ *is a contour starting at the point* $\gamma - \infty$ *and terminating at the point* $\gamma + \infty$*, where* $\gamma \in \mathbb{R}$*.*

Bibliography

[Aar] Scott Aaronson. *P ?= NP*. URL: https://www.scottaaronson.com/papers/pnp.pdf.

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. ISBN: 9781139477369.

[Alt+03] Joseph B Altepeter et al. "Ancilla-assisted quantum process tomography". In: *Physical Review Letters* 90.19 (2003), p. 193601.

[Ara+13] Itai Arad et al. "An area law and sub-exponential algorithm for 1D systems". In: *arXiv:1301.1162 [cond-mat, physics:quant-ph]* (2013).

[Ara+17] Itai Arad et al. "Rigorous RG algorithms and area laws for low energy eigenstates in 1D". In: *Communications in Mathematical Physics* 356.1 (2017), pp. 65–105.

[ARR14] Ali Ahmed, Benjamin Recht, and Justin Romberg. "Blind deconvolution using convex programming". In: *IEEE Transactions on Information Theory* 60.3 (2014), pp. 1711–1732.

[AS09] Koenraad MR Audenaert and Stefan Scheel. "Quantum tomographic reconstruction with error bars: a Kalman filter approach". In: *New Journal of Physics* 11.2 (2009), p. 023028.

[Aud+08] Koenraad MR Audenaert et al. "Asymptotic error rates in quantum hypothesis testing". In: *Communications in Mathematical Physics* 279.1 (2008), pp. 251–283.

[Bau+13a] Tillmann Baumgratz et al. "A scalable maximum likelihood method for quantum state tomography". In: *New Journal of Physics* 15.12 (2013), p. 125004.

[Bau+13b] Tillmann Baumgratz et al. "Scalable reconstruction of density matrices". In: *Physical review letters* 111.2 (2013), p. 020401.

[BC16] William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016. ISBN: 9780470141151.

[BC17] Jacob C Bridgeman and Christopher T Chubb. "Hand-waving and interpretive dance: an introductory course on tensor networks". In: *Journal of Physics A: Mathematical and Theoretical* 50.22 (2017), p. 223001.

[BCG95] Lawrence D Brown, George Casella, and JT Gene Hwang. "Optimal confidence sets, bioequivalence, and the limacon of Pascal". In: *Journal of the American Statistical Association* 90.431 (1995), pp. 880–889.

[Ben+17] Johann A Bengua et al. "Matrix product state for higher-order tensor compression and classification". In: *IEEE Transactions on Signal Processing* 65.15 (2017), pp. 4019–4030.

[BGW05] Arindam Banerjee, Xin Guo, and Hui Wang. "On the optimality of conditional expectation as a Bregman predictor". In: *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2664–2669.

[Bha13] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013. ISBN: 9780387948461.

[BK03] Mark S Byrd and Navin Khaneja. "Characterization of the positivity of the density matrix in terms of the coherence vector representation". In: *Physical Review A* 68.6 (2003), p. 062322.

[Blu12] Robin Blume-Kohout. "Robust error bars for quantum tomography". In: *arXiv preprint arXiv:1202.5270* (2012).

[BN98] Aharon Ben-Tal and Arkadi Nemirovski. "Robust convex optimization". In: *Mathematics of operations research* 23.4 (1998), pp. 769–805.

[BPT15] Johann A Bengua, Ho N Phien, and Hoang D Tuan. "Optimal feature extraction and classification of tensors via matrix product state decomposition". In: *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE. 2015, pp. 669–672.

[BR13a] Quentin Berthet and Philippe Rigollet. "Complexity theoretic lower bounds for sparse principal component detection". In: *Conference on Learning Theory*. 2013, pp. 1046–1066.

[BR13b] Quentin Berthet and Philippe Rigollet. "Computational lower bounds for sparse PCA". In: *arXiv preprint arXiv:1304.0828* (2013).

[Bro12] Andrew Browder. *Mathematical analysis: an introduction*. Springer Science & Business Media, 2012. ISBN: 9781461207153.

[Buž+98] V Bužek et al. "Reconstruction of quantum states of spin systems: From quantum Bayesian inference to quantum tomography". In: *Annals of Physics* 266.2 (1998), pp. 454–496.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. ISBN: 9780521833783.

[Car+15a] Jacques Carolan et al. "Universal linear optics". In: *Science* 349.6249 (2015), pp. 711–716.

[Car+15b] Alexandra Carpentier et al. “Uncertainty quantification for matrix compressed sensing and quantum tomography problems”. In: *arXiv preprint arXiv:1504.03234* (2015).

[CB02] George Casella and Roger L Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002. ISBN: 9780495391876.

[Che+17] Zhongming Chen et al. “Parallelized tensor train learning of polynomial classifiers”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2017).

[Che15] Yudong Chen. “Incoherence-Optimal Matrix Completion.” In: *IEEE Trans. Information Theory* 61.5 (2015), pp. 2909–2923.

[Chi+06] L Childress et al. “Coherent dynamics of coupled electron and nuclear spin qubits in diamond”. In: *Science* 314.5797 (2006), pp. 281–285.

[CL14] Emmanuel J Candès and Xiaodong Li. “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns”. In: *Foundations of Computational Mathematics* 14.5 (2014), pp. 1017–1026.

[CMR06] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models*. Springer Science & Business Media, 2006. ISBN: 9780387289823.

[Con18] Emily Conover. *Google Moves toward Quantum Supremacy with 72-Qubit Computer*. Mar. 5, 2018. URL: `https://www.sciencenews.org/article/google-moves-toward-quantum-supremacy-72-qubit-computer` (visited on 03/06/2018).

[Coo06] Stephen Cook. “The P versus NP problem”. In: *The millennium prize problems* (2006), p. 86.

[Coo90] Gregory F Cooper. “The computational complexity of probabilistic inference using Bayesian belief networks”. In: *Artificial intelligence* 42.2-3 (1990), pp. 393–405.

[CP11] Emmanuel J Candes and Yaniv Plan. “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements”. In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2342–2359.

[CR09] Emmanuel J Candès and Benjamin Recht. “Exact matrix completion via convex optimization”. In: *Foundations of Computational mathematics* 9.6 (2009), p. 717.

[CR12] Matthias Christandl and Renato Renner. “Reliable quantum state tomography”. In: *Physical Review Letters* 109.12 (2012), p. 120403.

[Cra+10] Marcus Cramer et al. “Efficient quantum state tomography”. In: *Nature communications* 1 (2010), p. 149.

[CRT06] Emmanuel J Candes, Justin K Romberg, and Terence Tao. “Stable signal recovery from incomplete and inaccurate measurements”. In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.8 (2006), pp. 1207–1223.

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming”. In: *Communications on Pure and Applied Mathematics* 66.8 (2013), pp. 1241–1274.

[CT10] Emmanuel J Candes and Terence Tao. “The power of convex relaxation: Near-optimal matrix completion”. In: *IEEE Transactions on Information Theory* 56.5 (2010), pp. 2053–2080.

[CV14] Ben Cousins and Santosh Vempala. “A cubic algorithm for computing gaussian volume”. In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2014, pp. 1215–1228.

[CV15] Benjamin Cousins and Santosh Vempala. “Bypassing KLS: Gaussian Cooling and an O^*(n3) Volume Algorithm”. In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM. 2015, pp. 539–548.

[De +13] Gemma De las Cuevas et al. “Purifications of multipartite states: limitations and constructive methods”. In: *New Journal of Physics* 15.12 (2013), p. 123021.

[Der+97] R Derka et al. “From quantum Bayesian inference to quantum tomography”. In: *Journal of Fine Mechanics and Optics* 341 (1997), pp. 11–12.

[DH14] Laurent Demanet and Paul Hand. “Stable optimizationless recovery from phaseless linear measurements”. In: *Journal of Fourier Analysis and Applications* 20.1 (2014), pp. 199–221.

[Dha+16] Ish Dhand et al. “Accurate and precise characterization of linear optical interferometers”. In: *Journal of Optics* 18.3 (2016), p. 035204.

[Dha+18] I Dhand et al. “Proposal for Quantum Simulation via All-Optically-Generated Tensor Network States”. In: *Physical review letters* 120.13 (2018), p. 130501.

[DL08] Vin De Silva and Lek-Heng Lim. “Tensor rank and the ill-posedness of the best low-rank approximation problem”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1084–1127.

[DLR16] Sjoerd Dirksen, Guillaume Lecué, and Holger Rauhut. “On the gap between restricted isometry properties and sparse recovery conditions”. In: *IEEE Transactions on Information Theory* (2016).

[Don06] David L Donoho. “For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution”. In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.6 (2006), pp. 797–829.

[ECP10] Jens Eisert, Marcus Cramer, and Martin B Plenio. “Colloquium: Area laws for the entanglement entropy”. In: *Reviews of Modern Physics* 82.1 (2010), p. 277.

[EGS06] Michael J Evans, Irwin Guttman, and Tim Swartz. “Optimally and computations for relative surprise inferences”. In: *Canadian Journal of Statistics* 34.1 (2006), pp. 113–129.

[ENP12] Yonina C Eldar, Deanna Needell, and Yaniv Plan. “Uniqueness conditions for low-rank matrix recovery”. In: *Applied and Computational Harmonic Analysis* 33.2 (2012), pp. 309–314.

[ET94] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. ISBN: 9780412042317.

[Fan57] Ugo Fano. “Description of states in quantum mechanics by density matrix and operator techniques”. In: *Reviews of Modern Physics* 29.1 (1957), p. 74.

[FC98] Gary J Feldman and Robert D Cousins. “Unified approach to the classical statistical analysis of small signals”. In: *Physical Review D* 57.7 (1998), p. 3873.

[Fer14a] Christopher Ferrie. “High posterior density ellipsoids of quantum states”. In: *New Journal of Physics* 16.2 (2014), p. 023006.

[Fer14b] Christopher Ferrie. “Quantum model averaging”. In: *New Journal of Physics* 16.9 (2014), p. 093035.

[FL11] Steven T Flammia and Yi-Kai Liu. “Direct fidelity estimation from few Pauli measurements”. In: *Physical Review Letters* 106.23 (2011), p. 230501.

[Fla+12] Steven T Flammia et al. “Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators”. In: *New Journal of Physics* 14.9 (2012), p. 095022.

[FNW92] Mark Fannes, Bruno Nachtergaele, and Reinhard F Werner. "Finitely correlated states on quantum spin chains". In: *Communications in mathematical physics* 144.3 (1992), pp. 443–490.

[FR12] Charles W Fox and Stephen J Roberts. "A tutorial on variational Bayesian inference". In: *Artificial intelligence review* 38.2 (2012), pp. 85–95.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. New York, NY: Springer New York, 2013. ISBN: 9780817649470.

[FR16] Philippe Faist and Renato Renner. "Practical and reliable error bars in quantum tomography". In: *Physical review letters* 117.1 (2016), p. 010404.

[GE16] Yimin Ge and Jens Eisert. "Area laws and efficient descriptions of quantum many-body states". In: *New Journal of Physics* 18.8 (2016), p. 083026.

[Gel+95] Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995. ISBN: 9780412039911.

[GFF17] Christopher Granade, Christopher Ferrie, and Steven T Flammia. "Practical adaptive quantum tomography". In: *New Journal of Physics* 19.11 (2017), p. 113017.

[GJ79] Michael R Garey and David S Johnson. *Computers and intractability*. wh freeman New York, 1979. ISBN: 9780716710455.

[GKK15a] Lars Grasedyck, Melanie Kluge, and Sebastian Krämer. "Variants of alternating least squares tensor completion in the tensor train format". In: *SIAM Journal on Scientific Computing* 37.5 (2015), A2424–A2450.

[GKK15b] David Gross, Felix Krahmer, and Richard Kueng. "A partial derandomization of PhaseLift using spherical designs". In: *Journal of Fourier Analysis and Applications* 21.2 (2015), pp. 229–266.

[GKK17] David Gross, Felix Krahmer, and Richard Kueng. "Improved recovery guarantees for phase retrieval from coded diffraction patterns". In: *Applied and Computational Harmonic Analysis* 42.1 (2017), pp. 37–64.

[Gor88] Yehoram Gordon. "On Milman's inequality and random subspaces which escape through a mesh in R^n". In: *Geometric Aspects of Functional Analysis*. Springer, 1988, pp. 84–106.

[GPY17] Navid Ghadermarzy, Yaniv Plan, and Özgür Yılmaz. "Near-optimal sample complexity for convex tensor completion". In: *arXiv preprint arXiv:1711.04965* (2017).

[Gra10] Lars Grasedyck. “Hierarchical singular value decomposition of tensors”. In: *SIAM Journal on Matrix Analysis and Applications* 31.4 (2010), pp. 2029–2054.

[Gro+10] David Gross et al. “Quantum state tomography via compressed sensing”. In: *Physical review letters* 105.15 (2010), p. 150401.

[Gro11] D Gross. “Recovering Low-Rank Matrices From Few Coefficients in Any Basis”. In: *IEEE Transactions on Information Theory* 3.57 (2011), pp. 1548–1566.

[GST12] Amparo Gil, Javier Segura, and Nico M Temme. “Efficient and accurate algorithms for the computation and inversion of the incomplete gamma function ratios”. In: *SIAM Journal on Scientific Computing* 34.6 (2012), A2965–A2981.

[GT09] Otfried Gühne and Géza Tóth. “Entanglement detection”. In: *Physics Reports* 474.1-6 (2009), pp. 1–75.

[GV12] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012. ISBN: 9781421407943.

[Hac12] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*. Vol. 42. Springer Science & Business Media, 2012.

[Häf+05] Hartmut Häffner et al. “Scalable multiparticle entanglement of trapped ions”. In: *Nature* 438.7068 (2005), p. 643.

[Háj12] Alan Hájek. “Interpretations of Probability”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2012.

[Har+17] Nicholas C Harris et al. “Quantum transport simulations in a programmable nanophotonic processor”. In: *Nature Photonics* 11.7 (2017), p. 447.

[Has06] Matthew B Hastings. “Solving gapped Hamiltonians locally”. In: *Physical review b* 73.8 (2006), p. 085115.

[HC82] Jiunn Tzon Hwang and George Casella. “Minimax confidence sets for the mean of a multivariate normal distribution”. In: *The Annals of Statistics* (1982), pp. 868–881.

[HH12] Ferenc Huszár and Neil MT Houlsby. “Adaptive Bayesian quantum tomography”. In: *Physical Review A* 85.5 (2012), p. 052120.

[HJ94] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994. ISBN: 9780521467131.

[HL13] Christopher J Hillar and Lek-Heng Lim. “Most tensor problems are NP-hard”. In: *Journal of the ACM (JACM)* 60.6 (2013), p. 45.

[HMW13] Teiko Heinosaari, Luca Mazzarella, and Michael M Wolf. "Quantum tomography under prior information". In: *Communications in Mathematical Physics* 318.2 (2013), pp. 355–374.

[Hol+15] M Holzäpfel et al. "Scalable reconstruction of unitary processes and Hamiltonians". In: *Physical Review A* 91.4 (2015), p. 042129.

[HOM87] Chong-Ki Hong, Zhe-Yu Ou, and Leonard Mandel. "Measurement of subpicosecond time intervals between two photons by interference". In: *Physical review letters* 59.18 (1987), p. 2044.

[Hou+16] Zhibo Hou et al. "Full reconstruction of a 14-qubit state within four hours". In: *New Journal of Physics* 18.8 (2016), p. 083036.

[Hra+04] Zdenek Hradil et al. "Maximum-likelihood methods in quantum mechanics". In: vol. 649. Springer; 1999, 2004, pp. 59–112.

[HRS12] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. "The alternating linear scheme for tensor optimization in the tensor train format". In: *SIAM Journal on Scientific Computing* 34.2 (2012), A683–A713.

[JFH03] Miroslav Ježek, Jaromír Fiurášek, and Zdeněk Hradil. "Quantum inference of states and processes". In: *Physical Review A* 68.1 (2003), p. 012305.

[Jon91] KRW Jones. "Principles of quantum inference". In: *Annals of Physics* 207.1 (1991), pp. 140–170.

[Jos67] V. M. Joshi. "Inadmissibility of the Usual Confidence Sets for the Mean of a Multivariate Normal Population". In: *The Annals of Mathematical Statistics* 38.6 (1967), pp. 1868–1875.

[Jos69] VM Joshi. "Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population". In: *The Annals of Mathematical Statistics* 40.3 (1969), pp. 1042–1067.

[Kab+16] Maryia Kabanava et al. "Stable low-rank matrix recovery via null space properties". In: *Information and Inference: A Journal of the IMA* 5.4 (2016), pp. 405–441.

[KB09] Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". In: *SIAM review* 51.3 (2009), pp. 455–500.

[Kec16] Michael Kech. "Explicit frames for deterministic phase retrieval via phaselift". In: *Applied and Computational Harmonic Analysis* (2016).

[Kel+03] Hans Kellerer et al. "An efficient fully polynomial approximation scheme for the subset-sum problem". In: vol. 66. 2. Elsevier, 2003, pp. 349–370.

[Key07] John Maynard Keynes. *A Treatise on Probability*. Cosimo, Inc., 2007. ISBN: 9781602066960.

[KGE14] Martin Kliesch, David Gross, and Jens Eisert. "Matrix-product operators and states: NP-hardness and undecidability". In: *Physical review letters* 113.16 (2014), p. 160503.

[Kie12] Jack C. Kiefer. *Introduction to Statistical Inference*. Springer Science & Business Media, 2012. ISBN: 9781461395782.

[Kim03] Gen Kimura. "The Bloch vector for N-level systems". In: *Physics Letters A* 314.5-6 (2003), pp. 339–349.

[KKD15] Amir Kalev, Robert L Kosut, and Ivan H Deutsch. "Quantum tomography protocols with positivity are compressed sensing protocols". In: *NPJ Quantum Information* 1 (2015), p. 15018.

[KL18] Felix Krahmer and Yi-Kai Liu. "Phase retrieval without small-ball probability assumptions". In: *IEEE Transactions on Information Theory* 64.1 (2018), pp. 485–500.

[KM15] Vladimir Koltchinskii and Shahar Mendelson. "Bounding the smallest singular value of a random matrix without concentration". In: *International Mathematics Research Notices* 2015.23 (2015), pp. 12991–13008.

[Kni+15] Lukas Knips et al. "How long does it take to obtain a physical density matrix?" In: *arXiv preprint arXiv:1512.06866* (2015).

[KRB08] Michał Karpiński, Czesław Radzewicz, and Konrad Banaszek. "Fiber-optic realization of anisotropic depolarizing quantum channels". In: *JOSA B* 25.4 (2008), pp. 668–673.

[KRT17] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. "Low rank matrix recovery from rank one measurements". In: *Applied and Computational Harmonic Analysis* 42.1 (2017), pp. 88–116.

[KS13] Akshay Krishnamurthy and Aarti Singh. "Low-rank matrix and tensor completion via adaptive sampling". In: (2013), pp. 836–844.

[KSV14] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. "Low-rank tensor completion by Riemannian optimization". In: *BIT Numerical Mathematics* 54.2 (2014), pp. 447–468.

[KSZ91] A Klumper, A Schadschneider, and J Zittartz. "Equivalence and solution of anisotropic spin-1 models and generalized t-J fermion models in one dimension". In: *Journal of Physics A: Mathematical and General* 24.16 (1991), p. L955.

[KSZ92] A Klumper, A Schadschneider, and J Zittartz. "Groundstate properties of a generalized VBS-model". In: *Zeitschrift für Physik B Condensed Matter* 87.3 (1992), pp. 281–287.

[Kue15] Richard Kueng. "Low rank matrix recovery from few orthonormal basis measurements". In: *Sampling Theory and Applications (SampTA), 2015 International Conference on*. IEEE. 2015, pp. 402–406.

[KZG16] Richard Kueng, Huangjun Zhu, and David Gross. "Low rank matrix recovery from Clifford orbits". In: *arXiv preprint arXiv:1610.08070* (2016).

[Lan+17] BP Lanyon et al. "Efficient tomography of a quantum many-body system". In: *Nature Physics* 13.12 (2017), p. 1158.

[LC98] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 1998. ISBN: 9780387985022.

[Le 12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012. ISBN: 9781461249467.

[Li+16] Xikun Li et al. "Optimal error intervals for properties of the quantum state". In: *Physical Review A* 94.6 (2016), p. 062112.

[LL10] Nan Li and Baoxin Li. "Tensor completion for on-board compression of hyperspectral images". In: *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE. 2010, pp. 517–520.

[LLB16] Yanjun Li, Kiryung Lee, and Yoram Bresler. "Optimal sample complexity for stable matrix recovery". In: (2016), pp. 81–85.

[LO12] Anthony Laing and Jeremy L O'Brien. "Super-stable tomography of any linear optical device". In: *arXiv preprint arXiv:1208.2868* (2012).

[Lun+09] JS Lundeen et al. "Tomography of quantum detectors". In: *Nature Physics* 5.1 (2009), p. 27.

[Men14] Shahar Mendelson. "Learning without concentration". In: (2014), pp. 25–39.

[Mil15] David AB Miller. "Sorting out light". In: *Science* 347.6229 (2015), pp. 1423–1424.

[MN89] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. CRC Press, 1989. ISBN: 9780412317606.

[Mol+04] G Molina-Terriza et al. "Triggered qutrits for quantum communication protocols". In: *Physical review letters* 92.16 (2004), p. 167903.

[MS16] E Miles Stoudenmire and David J Schwab. "Supervised learning with quantum-inspired tensor networks". In: *arXiv preprint arXiv:1605.05775* (2016).

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. ISBN: 9780262018029.

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. ISBN: 9781139495486.

[NG+13] Richard Nickl, Sara van de Geer, et al. "Confidence sets in sparse regression". In: *The Annals of Statistics* 41.6 (2013), pp. 2852–2876.

[Nov+15] Alexander Novikov et al. "Tensorizing neural networks". In: (2015), pp. 442–450.

[OBr+04] Jeremy L O'Brien et al. "Quantum process tomography of a controlled-NOT gate". In: *Physical review letters* 93.8 (2004), p. 080502.

[Orú14] Román Orús. "A practical introduction to tensor networks: Matrix product states and projected entangled pair states". In: *Annals of Physics* 349 (2014), pp. 117–158.

[Ose11] Ivan V Oseledets. "Tensor-train decomposition". In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317.

[Ose18a] Ivan Oseledets. *The git repository for the TT-Toolbox*. 2018. URL: `https://github.com/oseledets/TT-Toolbox`.

[Ose18b] Ivan Oseledets. *ttpy: Python implementation of the TT-Toolbox*. 2018. URL: `https://github.com/oseledets/ttpy`.

[OW15] Ryan O'Donnell and John Wright. "Quantum spectrum testing". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM. 2015, pp. 529–538.

[Per+07] D. Perez-Garcia et al. "Matrix Product State Representations". In: *Quantum Info. Comput.* 7.5 (July 2007), pp. 401–430. ISSN: 1533-7146.

[Phi+16] Ho N Phien et al. "Efficient tensor completion: Low-rank tensor train". In: *arXiv preprint arXiv:1601.01083* (2016).

[PR04] Matteo Paris and Jaroslav Rehacek. *Quantum State Estimation*. Springer Science & Business Media, 2004. ISBN: 9783540223290.

[PS17] Aaron Potechin and David Steurer. "Exact tensor completion with sum-of-squares". In: *arXiv preprint arXiv:1702.06237* (2017).

[Qi+13] Bo Qi et al. "Quantum state tomography via linear regression estimation". In: *Scientific reports* 3 (2013), p. 3496.

[Rah+11] Saleh Rahimi-Keshari et al. "Quantum process tomography with coherent states". In: *New Journal of Physics* 13.1 (2011), p. 013006.

[Rah+13] Saleh Rahimi-Keshari et al. “Direct characterization of linear-optical networks”. In: *Optics express* 21.11 (2013), pp. 13450–13458.

[Rec+94] Michael Reck et al. “Experimental realization of any discrete unitary operator”. In: *Physical review letters* 73.1 (1994), p. 58.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization”. In: *SIAM review* 52.3 (2010), pp. 471–501.

[Rie+07] M Riebe et al. “Quantum teleportation with atoms: quantum process tomography”. In: *New Journal of Physics* 9.7 (2007), p. 211.

[Rip+08] Lars Rippe et al. “Experimental quantum-state tomography of a solid-state qubit”. In: *Physical Review A* 77.2 (2008), p. 022307.

[Rot96] Dan Roth. “On the hardness of approximate reasoning”. In: *Artificial Intelligence* 82.1-2 (1996), pp. 273–302.

[RSS15] Holger Rauhut, Reinhold Schneider, and Željka Stojanac. “Tensor completion in hierarchical tensor representations”. In: *Compressed Sensing and its Applications* (2015), pp. 419–450.

[RSS17] Holger Rauhut, Reinhold Schneider, and Željka Stojanac. “Low rank tensor recovery via iterative hard thresholding”. In: *Linear Algebra and its Applications* 523 (2017), pp. 220–262.

[RU13] Thorsten Rohwedder and André Uschmajew. “On local convergence of alternating schemes for optimization of convex problems in the tensor train format”. In: *SIAM Journal on Numerical Analysis* 51.2 (2013), pp. 1134–1162.

[Rud17] Terry Rudolph. “Why I am optimistic about the silicon-photonic route to quantum computing”. In: *APL Photonics* 2.3 (2017), p. 030901.

[SB18] Travis L Scholten and Robin Blume-Kohout. “Behavior of the maximum likelihood in quantum state tomography”. In: *New Journal of Physics* 20.2 (2018), p. 023050.

[SBC01] Ruediger Schack, Todd A Brun, and Carlton M Caves. “Quantum bayes rule”. In: *Physical Review A* 64.1 (2001), p. 014305.

[Sch+14] Christian Schwemmer et al. “Experimental comparison of efficient tomography schemes for a six-qubit state”. In: *Physical review letters* 113.4 (2014), p. 040503.

[Sch+15] Christian Schwemmer et al. “Systematic errors in current quantum state tomography tools”. In: *Physical review letters* 114.8 (2015), p. 080403.

[Sch+17a] Jochen Scheuer et al. “Robust techniques for polarization and detection of nuclear spin ensembles”. In: *Physical Review B* 96.17 (2017), p. 174436.

[Sch+17b] Ilai Schwartz et al. “Pulsed polarisation for robust DNP”. In: *arXiv preprint arXiv:1710.01508* (2017).

[Sch11] Ulrich Schollwöck. “The density-matrix renormalization group in the age of matrix product states”. In: *Annals of Physics* 326.1 (2011), pp. 96–192.

[Sch62] Samuel Schechter. “Iteration methods for nonlinear problems”. In: *Transactions of the American Mathematical Society* 104.1 (1962), pp. 179–189.

[Sco06] Andrew J Scott. “Tight informationally complete quantum measurements”. In: *Journal of Physics A: Mathematical and General* 39.43 (2006), p. 13507.

[Seo+16] Tae Joon Seok et al. “Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers”. In: *Optica* 3.1 (2016), pp. 64–70.

[Sev05] Thomas A. Severini. *Elements of Distribution Theory*. Cambridge University Press, 2005. ISBN: 9780521844727.

[SGS12] John A Smolin, Jay M Gambetta, and Graeme Smith. “Efficient method for computing the maximum-likelihood quantum state from measurements with additive gaussian noise”. In: *Physical review letters* 108.7 (2012), p. 070502.

[Sha+13] Jiangwei Shang et al. “Optimal error regions for quantum state estimation”. In: *New Journal of Physics* 15.12 (2013), p. 123026.

[She+17] Yichen Shen et al. “Deep learning with coherent nanophotonic circuits”. In: *Nature Photonics* 11.7 (2017), p. 441.

[Shi89] Nobuo Shinozaki. “Improved confidence sets for the mean of a multivariate normal distribution”. In: *Annals of the Institute of Statistical Mathematics* 41.2 (1989), pp. 331–346.

[Sil+16] Joshua W Silverstone et al. “Silicon quantum photonics”. In: *IEEE Journal of Selected Topics in Quantum Electronics* 22.6 (2016), pp. 390–402.

[Sla95] Paul B Slater. “Quantum coin-tossing in a Bayesian Jeffreys framework”. In: *Physics Letters A* 206.1-2 (1995), pp. 66–72.

[Spa+17] Nicolò Spagnolo et al. “Learning an unknown transformation via a genetic approach”. In: *Scientific Reports* 7.1 (2017), p. 14316.

[SRG+17] Daniel Suess, Łukasz Rudnicki, David Gross, et al. “Error regions in quantum state tomography: computational complexity caused by geometry of quantum states”. In: *New Journal of Physics* 19.9 (2017), p. 093013.

[SSK17a] Željka Stojanac, Daniel Suess, and Martin Kliesch. “On products of Gaussian random variables”. In: *arXiv preprint arXiv:1711.10516* (2017).

[SSK17b] Željka Stojanac, Daniel Suess, and Martin Kliesch. “On the distribution of a product of N Gaussian random variables”. In: *Wavelets and Sparsity XVII*. Vol. 10394. International Society for Optics and Photonics. 2017, p. 1039419.

[Ste+06] Matthias Steffen et al. “Measurement of the entanglement of two superconducting qubits via state tomography”. In: *Science* 313.5792 (2006), pp. 1423–1425.

[Ste+56] Charles Stein et al. “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California. 1956.

[Sto18] E. Miles Stoudenmire. *ITensor: A C++ library for rapidly creating efficient tensor network calculations*. 2018. URL: `https://github.com/ITensor/ITensor`.

[Sue17a] Daniel Suess. *mpnum: Matrix Product Representation library for Python*. 2017. URL: `https://github.com/dseuss/mpnum`.

[Sue17b] Daniel Suess. *pypllon: Characterising linear optical networks via Phaselift*. 2017. URL: `https://github.com/dseuss/pypllon`.

[SZN17] Jiangwei Shang, Zhengyun Zhang, and Hui Khoon Ng. “Superfast maximum-likelihood reconstruction for quantum tomography”. In: *Physical Review A* 95.6 (2017), p. 062336.

[Tai+15] Cheng Tai et al. “Convolutional neural networks with low-rank regularization”. In: *arXiv preprint arXiv:1511.06067* (2015).

[TB+97] Yu-Ling Tseng, Lawrence D Brown, et al. “Good exact confidence sets for a multivariate normal mean”. In: *The Annals of Statistics* 25.5 (1997), pp. 2228–2258.

[Tro15] Joel A Tropp. “Convex recovery of a structured signal from independent random linear measurements”. In: *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[TSW16] Max Tillmann, Christian Schmidt, and Philip Walther. “On unitary reconstruction of linear optical networks”. In: *Journal of Optics* 18.11 (2016), p. 114002.

[VC06] Frank Verstraete and J Ignacio Cirac. “Matrix product states represent ground states faithfully”. In: *Physical Review B* 73.9 (2006), p. 094423.

[Ver10] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027* (2010).

[VGC04] Frank Verstraete, Juan J Garcia-Ripoll, and Juan Ignacio Cirac. “Matrix product density operators: Simulation of finite-temperature and dissipative systems”. In: *Physical review letters* 93.20 (2004), p. 207204.

[VMC08] Frank Verstraete, Valentin Murg, and J Ignacio Cirac. “Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems”. In: *Advances in Physics* 57.2 (2008), pp. 143–224.

[Vor13] Vladislav Voroninski. “Quantum tomography from few full-rank observables”. In: *arXiv preprint arXiv:1309.7669* (2013).

[WAA16] Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. “Tensor completion by alternating minimization under the tensor train (TT) model”. In: *arXiv preprint arXiv:1609.05587* (2016).

[Wal63] Adriaan Walther. “The question of phase retrieval in optics”. In: *Optica Acta: International Journal of Optics* 10.1 (1963), pp. 41–49.

[Was13] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013. ISBN: 9781441923226.

[Whi92] Steven R White. “Density matrix formulation for quantum renormalization groups”. In: *Physical review letters* 69.19 (1992), p. 2863.

[Wie+15] Nathan Wiebe et al. “Bayesian inference via rejection filtering”. In: *arXiv preprint arXiv:1511.06458* (2015).

[WNH14] Hua Wang, Feiping Nie, and Heng Huang. “Low-rank Tensor Completion with Spatio-temporal Consistency”. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014, pp. 2846–2852.

[YH16] Yongxin Yang and Timothy Hospedales. “Deep multi-task representation learning: A tensor factorisation approach”. In: *arXiv preprint arXiv:1605.06391* (2016).

[ZH16] Xiaoyan Zhu and Ripei Hao. “Context-aware location recommendations with tensor factorization”. In: *Communications in China (ICCC), 2016 IEEE/CIC International Conference on*. IEEE. 2016, pp. 1–6.

[Zha16] Anru Zhang. “Cross: Efficient low-rank tensor completion”. In: *arXiv preprint arXiv:1611.01129* (2016).

[ZJD15] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. “Efficient matrix sensing using rank-1 gaussian measurements”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2015, pp. 3–18.

[ZV04] Michael Zwolak and Guifré Vidal. “Mixed-state dynamics in one-dimensional quantum lattice systems: a time-dependent superoperator renormalization algorithm”. In: *Physical review letters* 93.20 (2004), p. 207205.

[ZWJ14] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. “Lower bounds on the performance of polynomial-time algorithms for sparse linear regression”. In: (2014), pp. 921–948.

Teilpublikationen

- D. Suess, Ł. Rudnicki, T. O. Maciel, D. Gross: *Error regions in quantum state tomography: computational complexity caused by geometry of quantum states*, New J. Phys. 19 093013 (2017)
- Ž. Stojanac, D. Suess, M. Kliesch: *On the distribution of a product of N Gaussian random variables*, Proceedings Volume 10394, Wavelets and Sparsity XVII; 1039419 (2017)
- Ž. Stojanac, D. Suess, M. Kliesch, *On products of Gaussian random variables*, arXiv:1711.10516
- D. Suess, M. Holzaepfel, *mpnum: A matrix product representation library for Python*, Journal of Open Source Software, 2(20), 465 (2017)

Acknowledgments

First and foremost, I would like to thank David Gross for his continuous support, for the numerous opportunities he has provided for me, and for the tremendous freedom I enjoyed during my PhD.

I would also like to thank all former and current members of his group in Freiburg and Cologne. Particularly, I am deeply grateful to Mariela Boevska and Felipe Montealegre for the support they provided in organizing the submission of this thesis. I would like to thank Željka Stojanac, Johan Åberg, and Martin Kliesch for proof reading parts of this manuscript and providing crucial feedback. For his friendship, his advice inside and outside academia, and for providing me with an opportunity that changed my life fundamentally, I would like to thank Richard Kueng. Finally, I would like to thank my coauthors Milan Holzaepfel and Łukasz Rudnicki for productive collaboration and discussions.

I would like to thank Stephen Bartlett and his group in Sydney for the hospitality and enjoyable atmosphere during my stay there. I am also grateful to Marco Tomamichel and Chris Ferrie for the productive atmosphere during my visit.

I am deeply grateful to my parents for their unconditional support, love, and encouragement. I would like to thank my brother for his valuable advice.

I would like to thank Mercedeh Edrisi for keeping me grounded during the time of writing this thesis and giving me a good reason to finish it rather sooner than later.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von den angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. David Gross betreut worden.

Ort, Datum Unterschrift