Compression for quantum population coding

Yuxiang Yang, Ge Bai, Giulio Chiribella, and Masahito Hayashi *Fellow, IEEE*

Abstract

We study the compression of arbitrary parametric families of n identically prepared finite-dimensional quantum states, in a setting that can be regarded as a quantum analogue of population coding. For a family with f free parameters, we propose an asymptotically faithful protocol that requires a memory of overall size $(f/2)\log n$. Our construction uses a quantum version of local asymptotic normality and, as an intermediate step, solves the problem of the optimal compression of n identically prepared displaced thermal states. Our protocol achieves the ultimate bound predicted by quantum Shannon theory. In addition, we explore the minimum requirement for quantum memory: On the one hand, the amount of quantum memory used by our protocol can be made arbitrarily small compared to the overall memory cost; on the other hand, any protocol using only classical memory cannot be faithful.

Index Terms

Population coding, compression, quantum system, local asymptotic normality, identically prepared state

I. Introduction

Many problems in quantum information theory involve a source that prepares multiple copies of the same quantum state. This is the case, for example, of quantum tomography [1], quantum cloning [2], [3], and quantum state discrimination [4]. The state prepared by the source is generally unknown to the agent who has to carry out the task. Instead, the agent knows that the state belongs to some parametric family of density matrices $\{\rho_\theta\}_{\theta\in\Theta}$, with the parameter θ varying in the set Θ. Also, it is promised that all the particles emitted by the source are independently prepared in the same quantum state ρ_θ: when the source is used n times, it generates n quantum particles in the tensor product state $\rho_\theta^{\otimes n}$.

How much information is contained in the n-particle state $\rho_\theta^{\otimes n}$? One way to address this question is to quantify the minimum amount of memory needed to store the state. Solving this problem requires an optimization over all possible compression protocols. When the number of copies is large, it is tempting to use a classical protocol, wherein the parameter θ is estimated and the estimate is stored in a classical memory. However, this type of storage is generally not faithful, as shown in [5] for several examples of pure state families. In order to achieve a faithful storage, a non-zero amount of quantum memory is generally required. It is important to stress that the problem of storing the n-copy states $\{\rho_\theta^{\otimes n}\,,\,\theta\in\Theta\}$ in a quantum memory is different from the standard problem of quantum data compression [6], [7], [8]. In our scenario, the mixed state ρ_θ is not regarded as the average state of an information source, but, instead, as a physical encoding of the parameter θ. The goal of compression is to preserve the encoding of the parameter θ, by storing the state $\rho_\theta^{\otimes n}$ into a memory and retrieving it with high fidelity for all possible values of θ. To stress the difference with standard quantum compression, we refer to our scenario as *compression for quantum population coding*. The expression "quantum population coding" refers to the encoding of the parameter θ into the many-particle state $\rho_\theta^{\otimes n}$. We choose this expression in analogy with (classical) population coding, whereby a parameter is encoded into the population of n individuals [9]. The typical example of population coding arises in computational neuroscience, where the population consists of neurons and the parameter represents an external stimulus.

The compression for quantum population coding has been studied by Plesch and Bužek [10] in the case where ρ_θ is a pure qubit state and no error is tolerated (see also [11] for a prototype experimental implementation). A first extension to mixed states, higher dimensions, and non-zero error was proposed by some of us in [12]. The protocol therein was proven to be optimal under the assumption that the decoding operation must satisfy a suitable conservation law. Later, a new protocol that reaches the ultimate information-theoretic bound was found for qubit states [13]. The classical version of the problem was addressed in [14]. However, finding the optimal protocol for arbitrary parametric families of quantum states has remained as an open problem so far.

In this paper, we provide the general theory for the compression of n-tensor product state in a quantum parametric state family. We consider two categories of state families: families of finite-dimensional states and displaced thermal families of infinite-dimensional states. These two categories of state families turn out to be connected by the quantum version of local asymptotic normality (Q-LAN)[15], [16], [17], [18], which reduces n-tensor product of a finite-dimensional state locally to

The material in this paper will be presented in part at the 2017 IEEE International Symposium on Information Theory (ISIT 2017), Aachen (Germany), 25-30 June 2017.

Y. Yang (e-mail: yangyx09@connect.hku.hk) and G. Bai (e-mail: baig@connect.hku.hk) are with the Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

G. Chiribella is with the Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong, and Canadian Institute for Advanced Research, CIFAR Program in Quantum Information Science (e-mail: giulio@cs.hku.hk)

M. Hayashi is with the Graduate School of Mathematics, Nagoya University, Furocho, Chikusaku, Nagoya, 464-860, Japan, and Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117542. (e-mail: masahito@math.nagoya-u.ac.jp)

a displaced thermal state. As the first step, we discuss this kind of compression for the thermal states family, which can be regarded as the quantum extension of the Gaussian distribution. In the next step, employing Q-LAN, we reduce the problem of compressing generic finite-dimensional states to the case of displaced thermal states. Unlike previous works, our protocol does not require any assumption on the symmetry of the state family. In addition, an intriguing feature of this protocol is that the ratio between the size of quantum memory and the size of classical memory can be made arbitrarily close to but not equal to zero. Such a feature is not an incidence but an essence: for identically prepared displaced thermal states and qudit states, we show that any compression protocol using only classical memory must have non-vanishing error.

The rest of the paper is structured as follows. In Section III we study the compression of displaced thermal states. In Section IV we propose a protocol for the compression of identically prepared finite-dimensional states. Optimality of the protocols is proven later in Section VI. In Section V we show that it is necessary to use quantum memory to achieve faithful compression. Finally, we conclude the paper by some discussions in Section VII.

II. Main result.

The main result of our work is the optimal compression of identically prepared quantum states. We consider two major categories of states: finite dimensional (i.e. qudit) states and displaced thermal states. The key question addressed here is what are the minimal amounts of memory required to encode these families of states, in a way that they can be recovered with an error vanishing in the number of input copies. The memory cost essentially depends on the (sub)family from which the states are drawn. For instance, the memory cost for diagonal qudit states (i.e. classical probability distributions) should be less than the cost for general qudit states. As a consequence, we need to specify the state subfamilies being considered before stating the main result.

We begin by introducing the parameterization for $d(<\infty)$-dimensional non-degenerate quantum systems, whose states can be generated by rotating a fixed state $\rho_0(\mu)$ with spectrum μ, i.e.

$$\rho_\theta = U_\xi \rho_0(\mu) U_\xi^\dagger \tag{1}$$

where

$$U_\xi = \exp\left[i \left(\sum_{1\le j<k\le d} \frac{\xi^R_{j,k} T_{j,k} + \xi^I_{j,k} T_{k,j}}{\sqrt{\mu_j - \mu_k}} \right) \right] \tag{2}$$

$$T_{j,k} = iE_{j,k} - iE_{k,j} \qquad T_{k,j} = E_{j,k} + E_{k,j} \tag{3}$$

is the exponential form of a $\mathsf{SU}(d)$ element, $E_{j,k}$ is a $d\times d$ matrix with entry (j,k) equal to 1 and other entries equal to 0. Therefore, any non-degenerate qudit state can be parameterized as ρ_θ, where $\theta=(\mu,\xi)\in\mathbb{R}^{d^2-1}$ is a vector of parameters with $\mu=(\mu_1,\mu_2,\dots,\mu_{d-1})$ being its spectrum, ordered as $\mu_1>\dots>\mu_{d-1}>0$, and $\xi=(\dots,\xi^R_{j,k},\xi^I_{j,k},\dots)(1\le j<k\le d)$ being the rotation parameters. A qudit state (sub)family will be denoted as $\{\rho_\theta^{\otimes n}\}_{\theta\in\Theta}$ with $\Theta=\Theta_1\times\Theta_2\times\dots\times\Theta_{d^2-1}$ where Θ_i is the range of the i-th component of θ. For simplicity of discussion, we assume $\Theta_i=[a_i,a_i']$ for $a_i,a_i'\in\mathbb{R}$. We call a component of μ a free *classical parameter* if its range is $[\mu_i,\mu_i']$ for $\mu_i'>\mu_i$ and a component of ξ a free *quantum parameter* if its range is $[\xi_i,\xi_i']$ for $\xi_i'>\xi_i$. Otherwise the range set of the parameter contains only one value and the parameter is fixed. We denote by f_c (f_q) the number of free classical (quantum) parameters of the (sub)family.

Next we introduce displaced thermal states, which are a type of infinite-dimensional states frequently encountered in quantum optics. Displaced thermal states can be regarded as coherent states subject to Gaussian additive noise. A coherent state, describing the state of a laser under ideal conditions, is parametrized as

$$|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}} \sum_{k=0}^{\infty} \frac{\alpha^k}{\sqrt{k!}} |k\rangle \qquad \alpha = |\alpha| e^{iT} \quad T\in[0,2\pi)\,,$$

where $\{|k\rangle\}$ is the photon number basis. Gaussian additive noise is described by the quantum channel (completely positive trace preserving map)

$$\mathcal{N}_\beta(\rho) = \int \mathrm{d}^2\mu \left(\frac{1-\beta}{\pi\beta}\right) e^{-\frac{(1-\beta)|\mu|^2}{\beta}} D_\mu \rho D_\mu^\dagger \tag{4}$$

where $D_\mu=\exp(\mu\hat{a}^\dagger-\bar{\mu}\hat{a})$ is the displacement operator and $\beta\in[0,1)$ is a suitable parameter. When applied to an input coherent state, the Gaussian additive noise outputs the displaced thermal state

$$\rho_{\alpha,\beta} = D_\alpha \rho_\beta^{(\mathrm{thm})} D_\alpha^\dagger \qquad \alpha=|\alpha|e^{iT} \quad T\in[0,2\pi), \tag{5}$$

with

$$\rho_\beta^{(\mathrm{thm})} = \mathcal{N}_\beta(|0\rangle\langle 0|) = (1-\beta)\sum_{i=0}^{\infty} \beta^i |i\rangle\langle i|\,. \tag{6}$$

We consider a (sub)family of n identically prepared displaced thermal states, denoted by $\{\rho_{\alpha,\beta}^{\otimes n}\}_{(\beta,\alpha)\in\Theta}$ with $\Theta=\Theta_\beta\times\Theta_\alpha$ being the parameter space. Similar as the qudit state case, there are three real parameters for the displaced thermal state family: the thermal parameter β, the strength of the displacement $|\alpha|$, and the phase $T=\arg\alpha$. β is a classical parameter, specifying the probability distribution of the eigenvalues, while $|\alpha|$ and T are quantum parameters, determining the eigenstates of the density matrix. Each of the three parameters is free if its range is $[a,b]$ for $a<b$, otherwise we assume that the range set contains only a single value and the parameter is fixed. Again, we denote by f_c (f_q) the number of free classical (quantum) parameters of the (sub)family.

A compression protocol consists of two components: the encoder, which compresses the input state into a memory, and the decoder, which recovers the state from the memory. A protocol is thus represented by a couple of quantum channels (completely positive trace-preserving linear maps) $(\mathcal{E},\mathcal{D})$ characterizing the encoder and the decoder, respectively. We focus on *faithful* compression protocols, whose error vanishes in the large n limit. As a measure of error, we choose the supremum of the trace distance between the original state and the recovered state $\mathcal{D}\circ\mathcal{E}(\rho_\theta^{\otimes n})$

$$\epsilon:=\sup_{\theta\in\Theta}\frac{1}{2}\|\rho_\theta^{\otimes n}-\mathcal{D}\circ\mathcal{E}(\rho_\theta^{\otimes n})\|_1\,. \tag{7}$$

The main result of our work is the following:

Theorem 1. *Let $\{\rho_\theta^{\otimes n}\}_{\theta=(\mu,\xi)\in\Theta}$ be the state (sub)family of n identical displaced thermal states or non-degenerate qudit states with f_c free classical parameters and f_q free quantum parameters. For any $\delta\in(0,2/9)$, the state family can be compressed into $[(1/2+\delta)f_c+(1/2)f_q]\log n$ classical bits and $(f_q\delta)\log n$ qubits with an error $\epsilon=O\left(n^{-\delta/2}\right)+O\left(n^{-\kappa(\delta)}\right)$, where $\kappa(\delta)$ is the error of Q-LAN [18] given by Eq. (26). The compression is optimal, in the sense that any compression protocol requiring a memory of size $[(f_c+f_q)/2-\delta']\log n$ with $\delta'>0$ cannot be faithful.*

Theorem 1 states that to encode each free parameter a memory of size $(1/2+\delta)\log n$ is required. When the parameter is classical, the required memory is fully classical; when the parameter is quantum, a quantum memory of $\delta\log n$ qubits is required. Note that Theorem 1 solves the compression of several important (sub)families:

- The full model family of qudits ($f_c=d-1$ and $f_q=d(d-1)$).
- The classical qudit subfamily ($f_c=d-1$ and $f_q=0$) of d-dimensional classical probability distributions can be compressed into $(d/2)\log n$ classical bits, retrieving the result of [14].
- The phase-covariant qudit subfamily ($f_c=d-1$ and $f_q=d(d-1)/2$), where $\xi_{j,k}^I=0$ for any j, k.

III. Compression of displaced thermal states.

In this section, we consider the compression of identically prepared displaced thermal states $\rho_{\alpha,\beta}^{\otimes n}$ for 8 different subfamilies, corresponding to cases when β, $|\alpha|$ and T are either free or fixed. The memory costs are summarized in Table I, which matches the statement of Theorem 1. For instance, in Case 5 we have $f_q=1$ (T is a quantum parameter) and $f_c=1$ (β is a classical parameter), and therefore the memory costs are $\delta\log n$ qubits and $(1+\delta)\log n$ bits. On the other hand, the errors for all cases satisfy the unified bound

$$\epsilon=O\left(n^{-\delta/2}\right).$$

TABLE I
Compression rate for different state families. Here $\delta>0$ is an arbitrary positive constant.

Case	displacement $\alpha=\|\alpha\|e^{iT}$	thermal parameter β	quantum bits	classical bits
0	fixed	fixed	0	0
1	fixed	free	0	$(1/2+\delta)\log n$
2	free	fixed	$2\delta\log n$	$\log n$
3	free	free	$2\delta\log n$	$(3/2+\delta)\log n$
4	T free; $\|\alpha\|$ fixed	fixed	$\delta\log n$	$(1/2)\log n$
5	T free; $\|\alpha\|$ fixed	free	$\delta\log n$	$(1+\delta)\log n$
6	$\|\alpha\|$ free; T fixed	fixed	$\delta\log n$	$(1/2)\log n$
7	$\|\alpha\|$ free; T fixed	free	$\delta\log n$	$(1+\delta)\log n$

We now show details of the compression protocol for each case.

A. Case 1.

Since Case 0 (all parameters are fixed) is essentially trivial, we start from Case 1, where only β is a free parameter. An important property for n identically displaced thermal states is that they are equivalent, up to an unitary transformation, to a product of thermal states where the displacement appears only on one mode, namely

$$U_{\rm BS}\,\rho_{\alpha,\beta}^{\otimes n}\,U_{\rm BS}^\dagger=\rho_{\sqrt{n}\alpha,\beta}\otimes\left(\rho_\beta^{(\rm thm)}\right)^{\otimes(n-1)}\,, \tag{8}$$

where $U_{\rm BS}$ is a unitary operator, which can be physically implemented through a suitable arrangement of beam splitters [19], [20]. Using this property, we can construct a protocol that separately compresses the displaced thermal state $\rho_{\sqrt{n}\alpha,\beta}$ and the $n-1$ thermal modes. For the compression of the thermal modes, we have the following lemma.

Lemma 1 (Compression of identically prepared thermal states)**.** *For any $x>0$, there exists a protocol $\left(\mathcal{E}^{(\rm thm)}_{n,x},\mathcal{D}^{(\rm thm)}_{n,x}\right)$ compressing n copies of a thermal state $\rho^{(\rm thm)}_{\beta}$ into $(1/2+x)\log n$ classical bits with error*

$$\epsilon_{\rm thm}=O\left(n^{-x}\right). \tag{9}$$

The proof of the above lemma can be found in Appendix. We emphasize that no quantum memory is required to encode thermal states. For any $\delta>0$, we can then construct the compression protocol as follows:

- *Encoder.*
  1) First perform on each input copy the displacement operation $\mathcal{D}_{-\alpha}$, defined by
  $$\mathcal{D}_{\mu}(\cdot):=D_{\mu}\cdot D_{\mu}^{\dagger}$$
  where $D_{\mu}=\exp(\mu\hat{a}^{\dagger}-\bar{\mu}\hat{a})$ is the displacement operator. The displacement operation transforms each input copy $\rho_{\alpha,\beta}$ into the thermal state $\rho^{(\rm thm)}_{\beta}$.
  2) Then apply the thermal state encoder $\mathcal{E}^{(\rm thm)}_{n,\delta}$ on the n-mode state and the outcome is encoded in a classical memory.
- *Decoder.*
  1) First use the thermal state decoder $\mathcal{D}^{(\rm thm)}_{n,\delta}$ to recover the n copies of the thermal state $\rho^{(\rm thm)}_{\beta}$ from the classical memory.
  2) Perform the displacement operation $\mathcal{D}_{\alpha}$ on each mode.

Note that the input state can be regarded as the state of n optical modes with each mode in a displaced thermal state, and thus the compression protocol can be regarded as a sequence of operations on the n-mode system. We shall use the term "mode" throughout this section.

The above protocol requires only $(1/2+\delta)\log n$ bits for any $\delta>0$ by Lemma 1. The error is also the same as in Lemma 1, for the displacement operations are unitary.

B. Case 2.

Next we study the more involved case when the displacement α is free. Let us first look at the idea how to compress before going into details of the compression protocol. To save as much quantum memory as possible, we may consider to estimate α and store the outcome in a classical memory. However, any estimate of α comes at the price of disturbing the input state, and, as shown later in Section V, the distortion caused by measurements on all input copies is so large that the state cannot be recovered faithfully. Instead of full measurements, here we adopt a strategy of measuring only a small portion of the input copies so that we still get an estimate of α, while the distortion can still be recovered. We then perform coherent quantum protocols based on the information gained by the estimation. Such an idea of gently testing the input state and then performing coherent compression is key to this work, which allows us to convert the memory cost from quantum to classical as much as possible.

For any $\delta>0$, the protocol runs as follows (see also Fig 1 for a flowchart illustration):

- *Preprocessing.* A preprocessing procedure is needed in order to store the estimate of α: Divide the range of α into n zones, each labeled by a point $\hat{\alpha}_i$ in it, so that $|\alpha'-\alpha''|=O(n^{-1/2})$ (note that α is complex) for any α',α'' in the same zone.
- *Encoder.*
  1) First perform the unitary channel $\mathcal{U}_{\rm BS}(\cdot)=U_{\rm BS}\cdot U_{\rm BS}^{\dagger}$ on the input state, where $U_{\rm BS}$ is the unitary defined by Eq. (8). The output state has the form $\rho_{\sqrt{n}\alpha,\beta}\otimes\left(\rho^{(\rm thm)}_{\beta}\right)^{\otimes(n-1)}$.
  2) Send the first and the last mode through a beam splitter with transitivity $n^{-\delta}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta}}\alpha,\beta}\otimes\left(\rho^{(\rm thm)}_{\beta}\right)^{\otimes(n-2)}\otimes\rho_{\sqrt{n^{1-\delta}}\alpha,\beta}$.
  3) Estimate α by the heterodyne measurement $\{{\rm d}^2\,\alpha'|\alpha'\rangle\langle\alpha'|\}$ on the last mode, which yields an estimate $\hat{\alpha}=\alpha'/\sqrt{n^{1-\delta}}$ with the probability distribution
  $$Q(\hat{\alpha}|\alpha)=(1-\beta)\exp[-(1-\beta)n^{1-\delta}|\hat{\alpha}-\alpha|^2]. \tag{10}$$
  Encode the label of the zone $\hat{\alpha}^*$ containing $\hat{\alpha}$ in a classical memory.
  4) Displace the first mode with $\mathcal{D}_{-\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$.

5) The state of the first mode then goes through a truncation in the photon number basis, described by the channel $\mathcal{P}_{n^{2\delta}}$ defined as

$$\mathcal{P}_{n^{2\delta}}(\rho) = P_{n^{2\delta}}\rho P_{n^{2\delta}} + (1 - \mathrm{Tr}[P_{n^{2\delta}}\rho])|0\rangle\langle 0| \tag{11}$$

where

$$P_{n^{2\delta}} = \sum_{m=0}^{n^{2\delta}} |m\rangle\langle m|. \tag{12}$$

The output state on the first mode is encoded in a quantum memory.

- *Decoder.*
  1) First read $\hat{\alpha}^*$ and perform the displacement operation $\mathcal{D}_{\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$ on the state of the quantum memory.
  2) To recover the distortion caused by the estimation of α, a quantum amplifier [21] $\mathcal{A}^{\gamma_n}$ with $\gamma_n = 1/(1-n^{-\delta})$ is applied to the output state. A quantum amplifier is a device that increases the intensity of quantum light while preserving its phase information, defined by the following quantum channel

$$\mathcal{A}^{\gamma}(\rho) = \mathrm{Tr}_B\left[e^{\cosh^{-1}(\sqrt{\gamma})(a^\dagger b^\dagger - ab)}(\rho \otimes |0\rangle\langle 0|_B)e^{\cosh^{-1}(\sqrt{\gamma})(ab - a^\dagger b^\dagger)}\right] \tag{13}$$

  where a and b are the annihilation operators of the input mode and the ancillary mode B, respectively.
  3) Prepare the other $(n-1)$ modes in the thermal state $\rho_\beta^{(\mathrm{thm})}$.
  4) Perform on all the n modes the inverse of the unitary channel $\mathcal{U}_{\mathrm{BS}}$

$$\mathcal{U}_{\mathrm{BS}}^{-1}(\rho) = U_{\mathrm{BS}}^\dagger \rho U_{\mathrm{BS}}. \tag{14}$$

The memory cost consists of two parts: the cost of encoding the (rounded) value $\hat{\alpha}^*$ of the estimate which is $\log n$ bits and the cost of encoding the first mode (in a displaced thermal state) which is $2\delta\log n$ qubits. Overall, the protocol requires $2\delta\log n$ qubits and $\log n$ classical bits.

On the other hand, the error of the protocol can be split into three terms as

$$\begin{aligned}\epsilon \le &\int_{|\hat{\alpha}-\alpha|>n^{-1/2+3\delta/4}} Q(\mathrm{d}\,\hat{\alpha}|\alpha) \\ &+ \frac{1}{2}\sup_{\alpha,\beta}\sup_{\hat{\alpha}^*:|\hat{\alpha}^*-\alpha|\le n^{-1/2+3\delta/4}}\left\{\left\|\mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha,\beta}) - \rho_{\sqrt{n}\alpha,\beta}\right\|_1 + \left\|\mathcal{P}_{n^{2\delta}}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta}) - \rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta}\right\|_1\right\}.\end{aligned}$$

Now we bound them one by one. First, from Eq. (10) we get

$$\int_{|\hat{\alpha}-\alpha|>n^{-1/2+3\delta/4}} Q(\mathrm{d}\,\hat{\alpha}|\alpha) = \int_{|\hat{\alpha}-\alpha|>n^{-1/2+3\delta/4}} (1-\beta)\exp[-(1-\beta)n^{1-\delta}|\hat{\alpha}-\alpha|^2] = e^{-\Omega(n^{\delta/2})}.$$

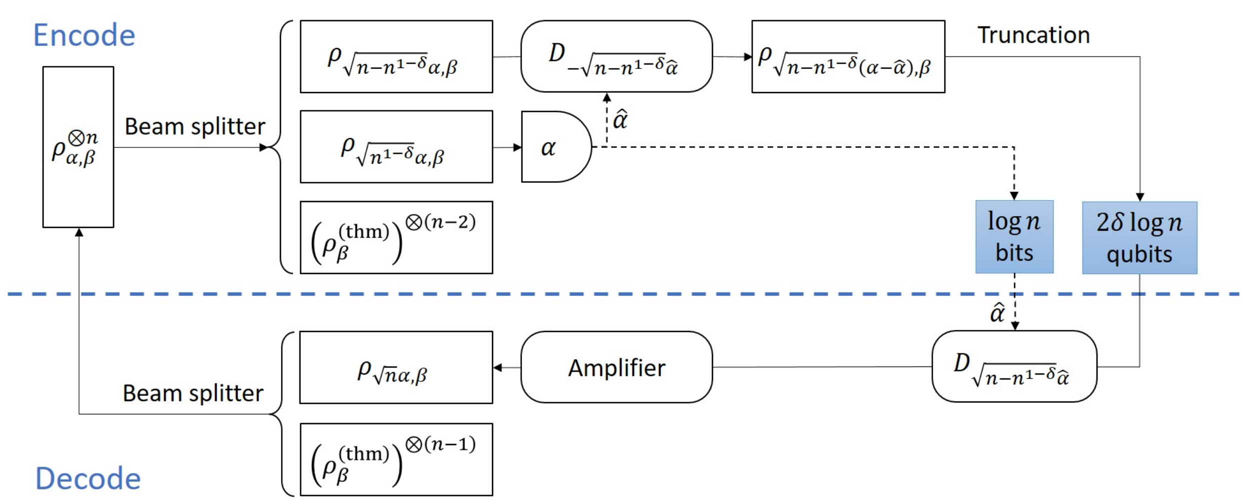

Fig. 1. **Compression protocol for fixed β and variable α.**

Second, explicit calculation shows that, when the input state is a displaced thermal state $\rho_{\alpha,\beta}$, the output state of the amplifier is

$$\mathcal{A}^{\gamma}\left(\rho_{\alpha,\beta}\right)=\rho_{\alpha,\beta'} \qquad \beta'=\frac{\beta\gamma+4(1-\beta)\gamma(\gamma-1)}{\gamma+(1-\beta)(\gamma-1)}. \tag{15}$$

We thus have

$$\frac{1}{2}\sup_{\alpha,\beta}\sup_{\hat{\alpha}^*:|\hat{\alpha}^*-\alpha|\le n^{-1/2+3\delta/4}}\left\|\mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha,\beta})-\rho_{\sqrt{n}\alpha,\beta}\right\|_1=O(n^{-\delta})$$

having used Eq. (19). Finally, the error of $\mathcal{P}_{n^{2\delta}}$ can be bounded using the following lemma, whose proof can be found in Appendix.

Lemma 2 (Photon number truncation of displaced thermal states.)**.** *Define the channel* $\mathcal{P}_N$ *as*

$$\mathcal{P}_N(\rho)=P_N\rho P_N+(1-\operatorname{Tr}[P_N\rho])|0\rangle\langle 0| \tag{16}$$

where

$$P_N=\sum_{k=0}^{N}|k\rangle\langle k|. \tag{17}$$

When $N=\Omega\left(|\alpha|^{2+x}\right)$*,* $\mathcal{P}_N$ *satisfies*

$$\epsilon(\rho_{\alpha,\beta})=\frac{1}{2}\left\|\mathcal{P}_N\left(\rho_{\alpha,\beta}\right)-\rho_{\alpha,\beta}\right\|_1=\beta^{\Omega(N^{x/8})}+e^{-\Omega(N^{x/4})} \tag{18}$$

for any $0\le\beta<1$.

In our case we are using the projector $P_{n^{2\delta}}$ in Eq. (12), and the displacement is $\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*)$. Since $(n-n^{1-\delta})|\alpha-\hat{\alpha}^*|^2=O\left(n^{3\delta/2}\right)$, by Lemma 2 we obtain the bound

$$\frac{1}{2}\sup_{\alpha,\beta}\sup_{\hat{\alpha}^*:|\hat{\alpha}^*-\alpha|\le n^{-1/2+3\delta/4}}\left\|\mathcal{P}_{n^{2\delta}}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta})-\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta}\right\|_1=\beta^{\Omega\left(n^{\delta/12}\right)}+e^{-\Omega(n^{\delta/6})}$$

Summarizing the above bounds for each term of the error, we get

$$\epsilon=O\left(n^{-\delta}\right).$$

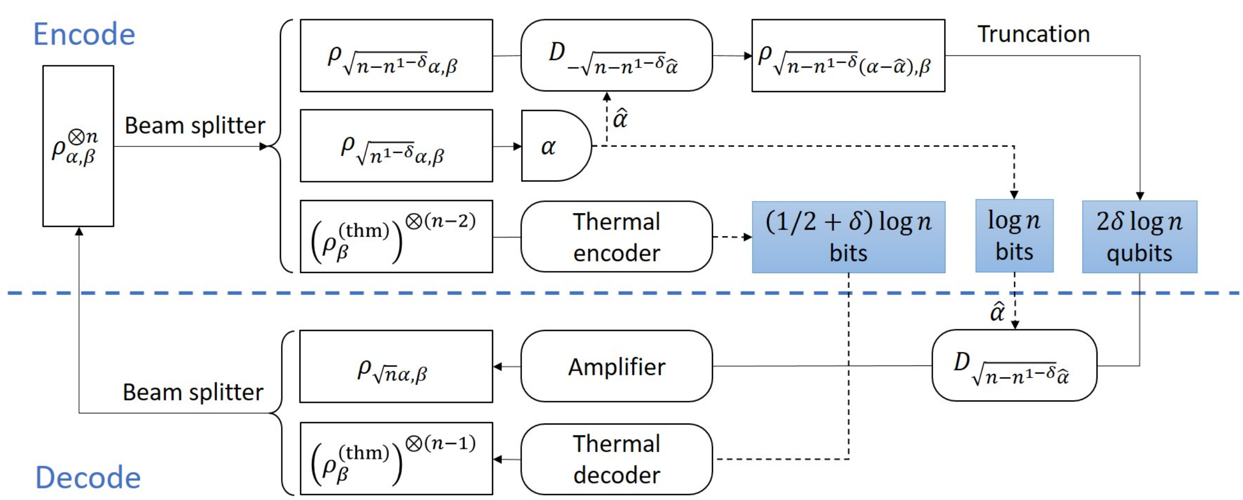

Fig. 2. **Compression protocol for variable α and β.**

C. Case 3.

Case 3 (free α, β) can be treated in a similar way as Case 2, except that we also have to estimate and encode β. Luckily, unlike α, the thermal parameter β can be estimated freely (i.e. without distortion of the input state), and thus its estimation strategy is simpler.

In detail, for any $\delta > 0$ we can then construct the protocol for displaced thermal states with free β, α as follows (see also Fig. 2 for a flowchart illustration):

- *Preprocessing.* Divide the range of α into n zones, each labeled by a point $\hat{\alpha}_i$ in it, so that $|\alpha' - \alpha''| = O(n^{-1/2})$ for any α', α'' in the same zone.
- *Encoder.*
  1) First perform the unitary channel $\mathcal{U}_{\rm BS}(\cdot) = U_{\rm BS} \cdot U_{\rm BS}^{\dagger}$ on the input state, where $U_{\rm BS}$ is the unitary defined by Eq. (8). The output state has the form $\rho_{\sqrt{n}\alpha,\beta} \otimes \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes(n-1)}$.
  2) Estimate β with the von Neumann measurement of the photon number on the $n-1$ copies of $\rho_{\beta}^{(\rm thm)}$ and denote by $\hat{\beta}$ the maximum likelihood estimate of β. Note that these copies will not be disturbed since they are diagonal in the photon number basis.
  3) Next, send the first and the last mode through a beam splitter with transitivity $n^{-\delta}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta}}\alpha,\beta} \otimes \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes(n-2)} \otimes \rho_{\sqrt{n^{1-\delta}}\alpha,\beta}$.
  4) Estimate α by the heterodyne measurement $\{{\rm d}^2\,\alpha' |\alpha'\rangle\langle\alpha'|\}$ on the last mode, which yields an estimate $\hat{\alpha} = \alpha'/\sqrt{n^{1-\delta}}$ with the probability distribution $Q(\hat{\alpha}|\alpha)$ (10). Encode the label of the zone $\hat{\alpha}^*$ containing $\hat{\alpha}$ in a classical memory.
  5) Displace the first mode with $\mathcal{D}_{-\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$.
  6) Prepare the n-th mode in the thermal state $\rho_{\hat{\beta}}^{(\rm thm)}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta} \otimes \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes(n-2)} \otimes \rho_{\hat{\beta}}^{(\rm thm)}$.
  7) The state of the first mode then goes through a truncation in the photon number basis, described by the channel $\mathcal{P}_{n^{2\delta}}$ defined by Eq. (16). The output state on the first mode is encoded in a quantum memory.
  8) Meanwhile, use the thermal state encoder $\mathcal{E}_{n-1,\delta}^{(\rm thm)}$ (see Lemma 1) to compress the remaining $n-1$ modes and encode the output state in a classical memory.
- *Decoder.*
  1) First read $\hat{\alpha}^*$ and perform the displacement $\mathcal{D}_{\sqrt{n-n^{1-\delta}}\hat{\alpha}^*}$ on the state of the quantum memory.
  2) Apply a quantum amplifier $\mathcal{A}^{\gamma_n}$ with $\gamma_n = 1/(1-n^{-\delta})$ to the state.
  3) Then use the thermal state decoder $\mathcal{D}_{n-1,\delta}^{(\rm thm)}$ to recover the other $(n-1)$ modes in the thermal state $\rho_{\beta}^{(\rm thm)}$ from the memory.
  4) Finally, perform the inverse of the unitary channel $\mathcal{U}_{\rm BS}$.

The memory cost consists of three parts: the cost of encoding the (rounded) value $\hat{\alpha}^*$ of the estimate which is $\log n$ bits, the cost of encoding the first mode (displaced thermal state) which is $2\delta \log n$ qubits, and the cost of encoding the other modes (thermal states) which is $(1/2+\delta)\log n$ bits. Overall, the protocol requires $2\delta\log n$ qubits and $(3/2+\delta)\log n$ classical bits.

On the other hand, the error of the protocol can be split into several terms as

$$\begin{aligned}\epsilon \leq\ & \frac{1}{2}\sup_{\beta}\left\|\mathcal{D}_{n,\delta}^{(\rm thm)} \circ \mathcal{E}_{n,\delta}^{(\rm thm)}\left[\int {\rm d}\,\hat{\beta}\, P({\rm d}\,\hat{\beta}|\beta)\, \rho_{\hat{\beta}}^{(\rm thm)} \otimes \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes(n-1)}\right] - \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes n}\right\|_1 \\ &+ \frac{1}{2}\sup_{\alpha,\beta}\sup_{\hat{\alpha}^*:|\hat{\alpha}^*-\alpha|\leq n^{-1/2+3\delta/4}}\left\{\left\|\mathcal{P}_N^{(\rm num)}(\rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta}) - \rho_{\sqrt{n-n^{1-\delta}}(\alpha-\hat{\alpha}^*),\beta}\right\|_1 + \left\|\mathcal{A}^{\gamma_n}(\rho_{\sqrt{n-n^{1-\delta}}\alpha,\beta}) - \rho_{\sqrt{n}\alpha,\beta}\right\|_1\right\} \\ &+ \int_{|\hat{\alpha}-\alpha|>n^{-1/2+3\delta/4}} Q({\rm d}\,\hat{\alpha}|\alpha)\end{aligned}$$

where $P({\rm d}\,\hat{\beta}|\beta)$ denotes the probability distribution of the estimate for the input value β. The latter three terms are the errors of estimation and compression of the displaced thermal state $\rho_{\sqrt{n}\alpha,\beta}$ of the first mode, which have been bounded in the previous subsection as $O(n^{-\delta})$. We now bound the first term ϵ_β, which is the error of compressing and estimating thermal states. ϵ_β can be further split into two terms as

$$\epsilon_\beta \leq \frac{1}{2}\sup_{\beta}\left\{\left\|\mathcal{D}_{n,\delta}^{(\rm thm)} \circ \mathcal{E}_{n,\delta}^{(\rm thm)}\left[\left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes n}\right] - \left(\rho_{\beta}^{(\rm thm)}\right)^{\otimes n}\right\|_1 + \left\|\int P({\rm d}\,\hat{\beta}|\beta)\, \rho_{\hat{\beta}}^{(\rm thm)} - \rho_{\beta}^{(\rm thm)}\right\|_1\right\}.$$

On the right hand side of the inequality, the first error term is bounded by Lemma 1, while the second error term is bounded by the following property of the maximum likelihood estimate [22]

$$\int_{|\hat{\beta}-\beta|\geq l/\sqrt{nF_\beta}} P(\mathrm{d}\,\hat{\beta}|\beta) \leq \mathrm{erfc}\left(\frac{l}{\sqrt{2}}\right)$$

where $F_\beta = (\beta^2+1)/[\beta(1-\beta)^3]$ is the Fisher information of β and $\mathrm{erfc}(x) := (2/\pi)\int_x^\infty e^{-s^2}\,\mathrm{d}\,s$ is the complementary error function. ϵ_β can thus be bounded as

$$\begin{aligned}\epsilon_\beta &\leq O(n^{-\delta}) + \frac{1}{2}\sup_\beta \sup_{\hat{\beta}:|\hat{\beta}-\beta|\leq n^{-(1+\delta)/2}} \left\|\rho^{(\mathrm{thm})}_{\hat{\beta}} - \rho^{(\mathrm{thm})}_{\beta}\right\|_1 + e^{-\Omega(n^\delta)}\\ &= O(n^{-\delta})\end{aligned}$$

having used the tail property of complementary error function and the property

$$\left\|\rho^{(\mathrm{thm})}_{\alpha} - \rho^{(\mathrm{thm})}_{\beta}\right\|_1 = \frac{2|\alpha-\beta|}{(1-\alpha)^2} + O(|\alpha-\beta|^2). \quad (19)$$

Summarizing the above bounds for each term of the error, we get

$$\epsilon = O\left(n^{-\delta}\right).$$

D. Case 4 and Case 6.

In Case 4 (free T, fixed $|\alpha|$, β) and Case 6 (free $|\alpha|$, fixed T, β), the displacement α is partially known. Such a knowledge allows us to reduce the amount of memory. Since the protocols for these two cases are very similar. The protocol for Case 4 runs as in the following:

- *Preprocessing.* Divide the range of T into $n^{-1/2}$ zones, each labeled by a point $\hat{T}_i$ in it, so that $|T'-T''| = O(n^{-1/2})$ for any T', T'' in the same zone.
- *Encoder.*
  1) First perform the unitary channel $\mathcal{U}_{\mathrm{BS}}(\cdot)$ on the input state to transform it into $\rho_{\sqrt{n}\alpha,\beta}\otimes\left(\rho^{(\mathrm{thm})}_\beta\right)^{\otimes(n-1)}$.
  2) Next, send the first and the last mode through a beam splitter with transitivity $n^{-\delta/2}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta/2}}\alpha,\beta}\otimes\left(\rho^{(\mathrm{thm})}_\beta\right)^{\otimes(n-2)}\otimes\rho_{\sqrt{n^{1-\delta/2}}\alpha,\beta}$.
  3) Estimate T by the heterodyne measurement $\mathrm{d}^2\,\alpha'|\alpha'\rangle\langle\alpha'|$ on the last mode, which yields an estimate $\hat{T}$ which is the phase of α'. Encode the label of the zone $\hat{T}^*$ containing $\hat{T}$ in a classical memory.
  4) Displace the first mode with $\mathcal{D}_{-\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$ with $\hat{\alpha}^* := |\alpha|e^{i\hat{T}^*}$.
  5) Then send the state of the first mode through a truncation channel $\mathcal{P}_{n^\delta}$ defined in (16) and encode the output state in a quantum memory.
- *Decoder.*
  1) First read $\hat{\alpha}^*$ and perform the displacement $\mathcal{D}_{\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$ on the state of the quantum memory.
  2) Then apply the quantum amplifier $\mathcal{A}^{\gamma_n}$ with $\gamma_n = 1/(1-n^{-\delta/2})$.
  3) Prepare the other $(n-1)$ modes in the thermal state $\rho^{(\mathrm{thm})}_\beta$.
  4) Finally, perform $\mathcal{U}^{-1}_{\mathrm{BS}}$ on the output of $\mathcal{D}^{(\mathrm{thm})}_{n-1}$ and the quantum memory.

The protocol for Case 6 works in the same way except that $|\alpha|$ is estimated instead of T. For both cases the memory cost consists of three parts: the cost of encoding the (rounded) value $\hat{\alpha}^*$ of the estimate which is $(1/2)\log n$ bits, the cost of encoding the first mode (displaced thermal state) which is $\delta\log n$ qubits, and the cost of encoding the other modes (thermal states) which is $(1/2+\delta)\log n$ bits. Overall, the protocol requires $\delta\log n$ qubits and $(1+2\delta)\log n$ classical bits. The error can be bounded as previous as

$$\epsilon = O\left(n^{-\delta/2}\right).$$

E. Case 5 and Case 7.

Case 5 (fixed $|\alpha|$, free T, β) and Case 7 (fixed T, free $|\alpha|$, β) can be treated in the same way as Case 4 and Case 6, adding additional steps to estimate and encode β. The protocol for Case 5 runs as follows:

- *Preprocessing.* Divide the range of T into $n^{-1/2}$ zones, each labeled by a point $\hat{T}_i$ in it, so that $|T'-T''| = O(n^{-1/2})$ for any T', T'' in the same zone.
- *Encoder.*

1) First perform the unitary channel $\mathcal{U}_{\rm BS}(\cdot)$ on the input state to transform it into $\rho_{\sqrt{n}\alpha,\beta}\otimes\left(\rho^{(\rm thm)}_{\beta}\right)^{\otimes(n-1)}$.
2) Estimate β with the von Neumann measurement of the photon number on the $n-1$ copies of $\rho^{(\rm thm)}_{\beta}$. Denote by $\hat{\beta}$ the maximum likelihood estimate of β.
3) Next, send the first and the last mode through a beam splitter with transitivity $n^{-\delta/2}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta/2}}\alpha,\beta}\otimes\left(\rho^{(\rm thm)}_{\beta}\right)^{\otimes(n-2)}\otimes\rho_{\sqrt{n^{1-\delta/2}}\alpha,\beta}$.
4) Estimate T by the heterodyne measurement $\mathrm{d}^2\,\alpha'|\alpha'\rangle\langle\alpha'|$ on the last mode, which yields an estimate $\hat{T}$ which is the phase of α'. Encode the label of the zone $\hat{T}^*$ containing $\hat{T}$ in a classical memory.
5) Displace the first mode with $\mathcal{D}_{-\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$ with $\hat{\alpha}^*:=|\alpha|e^{i\hat{T}^*}$.
6) Prepare the n-th mode in the thermal state $\rho^{(\rm thm)}_{\hat{\beta}}$. The n-mode state is now $\rho_{\sqrt{n-n^{1-\delta/2}}(\alpha-\hat{\alpha}^*),\beta}\otimes\left(\rho^{(\rm thm)}_{\beta}\right)^{\otimes(n-2)}\otimes\rho^{(\rm thm)}_{\hat{\beta}}$.
7) Then send the state of the first mode through a truncation channel $\mathcal{P}_{n^\delta}$ defined in (16) and encode the output state in a quantum memory.
8) Meanwhile, use the thermal state encoder $\mathcal{E}^{(\rm thm)}_{n-1,\delta}$ (see Lemma 1) to compress the remaining $n-1$ modes and encode the output state in a classical memory.

- *Decoder.*
  1) First read $\hat{\alpha}^*$ and perform the displacement $\mathcal{D}_{\sqrt{n-n^{1-\delta/2}}\hat{\alpha}^*}$ on the state of the quantum memory.
  2) Then apply the quantum amplifier $\mathcal{A}^{\gamma_n}$ with $\gamma_n=1/(1-n^{-\delta/2})$.
  3) Meanwhile, use the thermal state decoder $\mathcal{D}^{(\rm thm)}_{n-1,\delta}$ to recover the other $(n-1)$ modes in the thermal state $\rho^{(\rm thm)}_{\beta}$ from the memory.
  4) Finally, perform $\mathcal{U}^{-1}_{\rm BS}$ on the output of $\mathcal{D}^{(\rm thm)}_{n-1}$ and the quantum memory.

The protocol for Case 7 works in the same way except that $|\alpha|$ is estimated instead of T. For both cases the memory cost consists of three parts: the cost of encoding the (rounded) value $\hat{\alpha}^*$ of the estimate which is $(1/2)\log n$ bits, the cost of encoding the first mode (displaced thermal state) which is $\delta\log n$ qubits, and the cost of encoding the other modes (thermal states) which is $(1/2+\delta)\log n$ bits. Overall, the protocol requires $\delta\log n$ qubits and $(1+2\delta)\log n$ classical bits. The error can be bounded as previous as

$$\epsilon=O\left(n^{-\delta/2}\right).$$

IV. Compression of identically prepared finite dimensional systems.

In this section, we study the compression of $d(<\infty)$-dimensional non-degenerate quantum systems, based on the results on displaced thermal states and quantum local asymptotic normality. We show that, just as for displaced thermal states, each free parameter of a qudit subfamily requires $(1/2+\delta)\log n$ memory for any $\delta>0$. Details of the compression protocol are introduced in the following.

A. The compression protocol

To construct a compression protocol, we need the following techniques:

- *Quantum local asymptotic normality (Q-LAN).* The quantum version of local asymptotic normality has been derived in several different contexts [16], [17], [18]. Here we use the result of [18], which states that n identical copies of a qudit state can be approximated by a classical-quantum Gaussian state in a sufficiently small neighborhood of a point $\theta_0\in\Theta$ for large n. Explicitly, for a fixed point $\theta_0=(\mu_0,\xi_0)$ and for a fixed $x\in(0,1)$, we define the following neighborhood

$$\Theta_{n,x}(\theta_0)=\left\{\theta=\theta_0+\delta\theta/\sqrt{n},\ |\ \|\delta\theta\|_\infty\le n^{\frac{x}{2}}\right\},\tag{20}$$

where $\|\delta\theta\|_\infty$ is the max vector norm defined as $\|\delta\theta\|_\infty:=\max_i(\delta\theta)_i$. Q-LAN states that every n-fold product state $\rho_\theta^{\otimes n}$ with θ in the neighborhood $\Theta_{n,x}(\theta_0)$ can be approximated by a classical-quantum Gaussian state. Specifically, the stated $\rho_\theta^{\otimes n}$ is approximated by the Gaussian state

$$G_{n,\theta}=N_{\delta\mu,I_{\mu_0}}\bigotimes_{j<k}\rho_{\alpha_{j,k},\beta_{j,k}},\tag{21}$$

where $N_{\delta\mu,I_{\mu_0}}$ is the multivariate normal distribution with mean $\delta\mu$ and covariance matrix I_{μ_0} (equal to the inverse of the quantum Fisher information of the eigenvalues μ, evaluated at μ_0) and $\rho_{\alpha_{j,k},\beta_{j,k}}$ is the displaced thermal state defined as

$$\rho_{\alpha_{j,k},\beta_{j,k}}=D_{\alpha_{j,k}}\rho^{(\rm thm)}_{\beta_{j,k}}D^\dagger_{\alpha_{j,k}}\tag{22}$$

$$\alpha_{j,k}=\frac{\delta\xi^R_{j,k}+i\delta\xi^I_{j,k}}{2\sqrt{(\mu_0)_j-(\mu_0)_k}}\qquad\beta_{j,k}=\frac{(\mu_0)_k}{(\mu_0)_j}.\tag{23}$$

where $(\mu_0)_j$ and $(\mu_0)_k$ are components of μ_0. The approximation is physically implemented by two quantum channels $\mathcal{T}_{\theta_0}^{(n)}$ and $\mathcal{S}_{\theta_0}^{(n)}$, satisfying the conditions

$$\sup_{\theta\in\Theta_{n,x}(\theta_0,c)}\left\|\mathcal{T}_{\theta_0}^{(n)}\left(\rho_\theta^{\otimes n}\right)-G_{n,\theta}\right\|_1=O\left(n^{-\kappa(x)}\right) \tag{24}$$

$$\sup_{\theta\in\Theta_{n,x}(\theta_0,c)}\left\|\rho_\theta^{\otimes n}-\mathcal{S}_{\theta_0}^{(n)}\left(G_{n,\theta}\right)\right\|_1=O\left(n^{-\kappa(x)}\right), \tag{25}$$

where $\|\cdot\|_1$ denotes the trace norm and

$$\kappa(x)=\min\left\{\frac{1-z-\eta}{2},\frac{1-x-2y}{2}-y,\frac{2-9\eta}{24}\right\} \tag{26}$$

under the constraints $1>z>1+x/2$, $y>0$, $\eta>0$ and $\eta>x-y$. We note that $\kappa(x)>0$ when $x\in[0,2/9)$.

- *Quantum state tomography.* State tomography is an important technique of quantum information processing which is used to determine the density matrix of an unknown quantum state. The role of tomography here is to provide a rough estimate of θ_0 so that we can apply Q-LAN. The tomography protocol used here [23] gives an estimate $\rho_{\hat{\theta}}$ of a qudit state ρ_θ with confidence

$$\mathbf{Prob}\left[\frac{1}{2}\|\rho_\theta-\rho_{\hat{\theta}}\|_1\leq\epsilon\right]\geq 1-(n+1)^{3d^2}e^{-n\epsilon^2} \tag{27}$$

using n copies of the state.

For any $\delta\in(0,2/9)$, our compression protocol runs as follows (see also Fig. 3 for a flowchart illustration):

- *Encode the input state into memories.* We break the encoding of $\rho_\theta^{\otimes n}$ into four steps:
  1) *Tomography.* First take out $n^{1-\delta/2}$ copies of ρ_θ for quantum tomography. In this way, one obtains a neighborhood $\Theta_{n,2\delta/3}(\theta_0)$ (note that $2\delta/3$ is not the only choice) that contains the state with confidence approaching one in the large n limit. To encode the outcome of the tomography, the parameter space Θ is discretized into a lattice of $n^{(f_c+f_q)/2}$ points, each point corresponding to an outcome of tomography. The point that is closest to θ_0 is encoded in a classical memory.
  2) *Q-LAN.* After the tomography step, $n-n^{1-\delta/2}$ copies remain for compression. Define

$$\gamma_n=1/(1-n^{-\delta/2}). \tag{28}$$

  so that the number of remaining copies is n/γ_n. The n/γ_n copies are sent through the channel $\mathcal{T}_{\theta_0}^{(n/\gamma_n)}$ (24) which outputs the Gaussian state $G_{(n/\gamma_n),\theta}$ defined by Eq. (21).

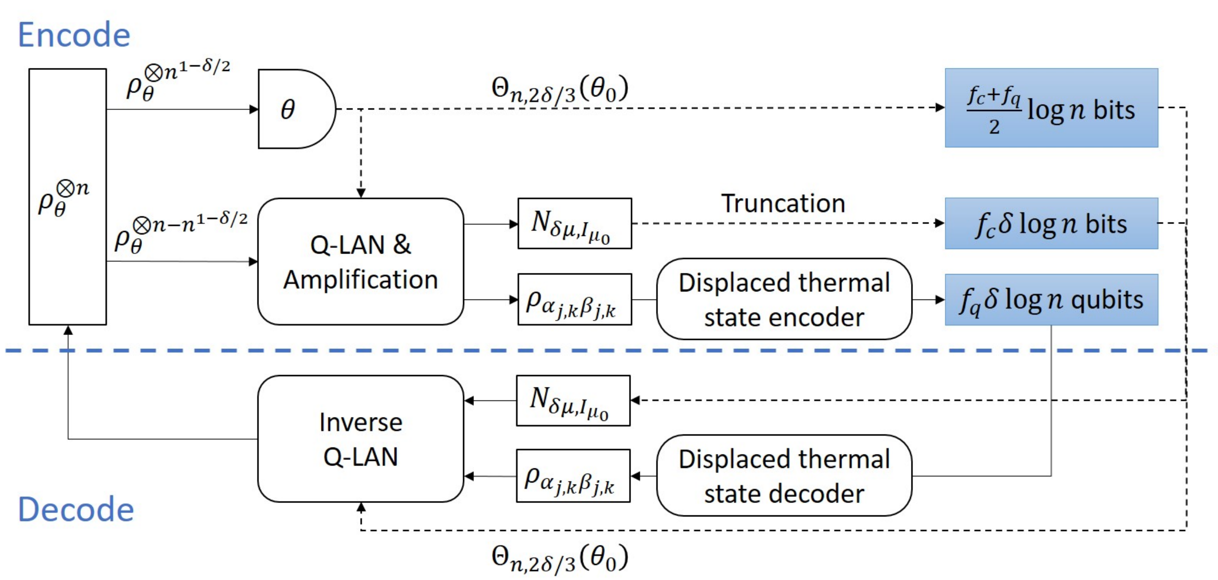

Fig. 3. **Compression protocol for qudit states.**

3) *Amplification.* Next, the Gaussian state is amplified to compensate the loss of input copies. The state $\rho_{\alpha_{j,k},\beta_{j,k}}$ of each quantum mode is amplified by $\mathcal{A}^{\gamma_n}$. The Gaussian distribution on the classical register is rescaled by a constant factor:
$$|u\rangle \to |\sqrt{\gamma_n}u\rangle,$$
where $\{|u\rangle\}$ stands for the Cartesian basis of the classical register. The whole amplification progress is described by the channel $\mathcal{A}_{\theta_0}^{(n/\gamma_n)\to n}$, whose action on any product state is
$$\mathcal{A}_{\theta_0}^{(n/\gamma_n)\to n}\left(\rho\otimes_{j<k}\sigma_{j,k}\right)=\sum_u\langle u|\rho|u\rangle|\sqrt{\gamma_n}u\rangle\langle\sqrt{\gamma_n}u|\bigotimes_{j<k}\mathcal{A}^{\gamma_n}(\sigma_{j,k}). \tag{29}$$
where u is summed over the basis of the classical register.

4) *Gaussian state compression.* Each quantum mode of the amplified Gaussian state is then truncated by $\mathcal{P}_{n^\delta}$ (16). The output state is then stored in a quantum memory. The state of the classical mode is compressed by an channel $\mathcal{P}_c$ that truncates the state into a $O(n^\delta)$-hypercube centered around the mean of the Gaussian. Explicitly, we have
$$\mathcal{P}_{\mathrm{c}}(\rho)=\sum_{\|u\|_\infty\le n^\delta}\langle u|\rho|u\rangle\,|u\rangle\langle u|+\left[1-\sum_{\|u'\|_\infty\le n^\delta}\langle u'|\rho|u'\rangle\right]|0\rangle\langle 0|. \tag{30}$$
The whole process is described by the channel
$$\mathcal{P}_{\theta_0}^{(n)}=\mathcal{P}_{\mathrm{c}}\bigotimes_{j<k}\mathcal{P}_{n^\delta}\,. \tag{31}$$

- *Recover the original state.* The state can be decompressed from the memory by sending the state of the hybrid memory through the channel $\mathcal{S}_{\theta_0}^{(n)}$ (25), which can be constructed by consulting the outcome of tomography.

B. Error analysis.

Here we bound the error of the protocol. Using the triangle inequality of trace distance, we split the overall error into four terms
$$\epsilon\le\epsilon_{\mathrm{tomo}}+\epsilon_{\mathrm{amp}}+\epsilon_{\mathrm{G}}+\epsilon_{\mathrm{LAN}}, \tag{32}$$
where
$$\epsilon_{\mathrm{tomo}}=\mathbf{Prob}\left[\theta\not\in\Theta_{n,2\delta/3}(\theta_0)\right] \tag{33}$$
$$\epsilon_{\mathrm{amp}}=\frac{1}{2}\sup_{\theta_0}\sup_{\theta\in\Theta_{n,2\delta/3}(\theta_0)}\left\|\mathcal{A}_{\theta_0}^{(n/\gamma_n)\to n}\left(G_{n/\gamma_n,\theta}\right)-G_{n,\theta}\right\|_1 \tag{34}$$
$$\epsilon_{\mathrm{G}}=\frac{1}{2}\sup_{\theta_0}\sup_{\theta\in\Theta_{n,2\delta/3}(\theta_0)}\left\|\mathcal{P}_{\theta_0}^{(n)}\left(G_{n,\theta}\right)-G_{n,\theta}\right\|_1 \tag{35}$$
$$\epsilon_{\mathrm{Q-LAN}}=\frac{1}{2}\sup_{\theta_0}\sup_{\theta\in\Theta_{n,2\delta/3}(\theta_0)}\left\{\left\|\mathcal{T}_{\theta_0}^{(n/\gamma_n)}\left(\rho_\theta^{\otimes(n/\gamma_n)}\right)-G_{n/\gamma_n,\theta}\right\|_1+\left\|\rho_\theta^{\otimes n}-\mathcal{S}_{\theta_0}^{(n)}\left(G_{n,\theta}\right)\right\|_1\right\} \tag{36}$$
are the error terms of tomography, amplification, truncation, and Q-LAN, respectively.

We first briefly review the process of tomography and then bound its error. Using $n^{1-\delta/2}$ copies of ρ_θ, an estimate $\hat\theta$ of θ is obtained. The estimate $\hat\theta$ is then encoded as a point in the lattice $\mathsf{L}:=\{\theta\in\Theta\mid|\theta_i|=z_i/(2\sqrt{n}),z_i\in\mathbb{N}\ \forall\ i\}$. Naturally, we encode the coordinate of the lattice point θ_0 that is closest to $\hat\theta$. Namely that we choose
$$\theta_0:=\operatorname*{argmin}_{\theta'\in\mathsf{L}}\|\hat\theta-\theta'\|_\infty. \tag{37}$$
Recall that $\|\theta\|_\infty:=\max_i(\theta)_i$. We now bound the error. By definition of the neighborhood $\Theta_{n,2\delta/3}(\theta_0)$ (20), we have
$$\epsilon_{\mathrm{tomo}}=\mathbf{Prob}\left[\|\theta-\theta_0\|_\infty>n^{-1/2+\delta/3}\right] \tag{38}$$
$$\le\mathbf{Prob}\left[\|\theta-\hat\theta\|_\infty+\|\hat\theta-\theta_0\|_\infty>n^{-1/2+\delta/3}\right] \tag{39}$$
$$\le\mathbf{Prob}\left[\|\theta-\hat\theta\|_\infty>n^{-1/2+\delta/3}(1-O(n^{-\delta/3}))\right]. \tag{40}$$
The first inequality comes from triangle inequality and the second inequality is derived by noticing $\min_{\theta\neq\theta'\in\mathsf{L}}\|\theta-\theta'\|_\infty=(1/2)n^{-1/2}$, which implies that $\|\hat\theta-\theta_0\|_\infty\le(1/2)n^{-1/2}$. To further bound the error, we notice that the parameter space

Θ allows for the Euclidean expansion of trace distance, namely that there exists a constant C such that $\|\rho_\theta - \rho_{\theta'}\|_1 = C\|\theta-\theta'\|_\infty + O(\|\theta-\theta'\|_\infty^2)$. And thus we have

$$\epsilon_{\text{tomo}} \le \textbf{Prob}\left[\frac{1}{2}\|\rho_\theta - \rho_{\hat\theta}\|_1 > (C/4)n^{-1/2+\delta/3}\right]. \tag{41}$$

Substituting ϵ with $(C/4)n^{-1/2+\delta/3}$ and n with $n^{1-\delta/2}$ in Eq. (27) we have

$$\epsilon_{\text{tomo}} = n^{-\Omega(n^{\delta/6})}. \tag{42}$$

Next we look at the error of amplification, which can be further split into two terms: the term of classical mode amplification and the term of quantum mode amplification. We first analyze the classical term, which comes from the rescaling of the Gaussian distribution. This operation shifts the center of the normal distribution from $\sqrt{\gamma_n^{-1}}\delta\mu$ (the classical part of $G_{n/\gamma_n,\theta}$) to $\delta\mu$, which coincides with that of $G_{n,\theta}$, while it also deforms the covariance matrix from I_{μ_0} to $\gamma_n I_{\mu_0}$. As a result, we consider the difference of the following distributions: $N_{\delta\mu,I_{\mu_0}}$ and $N_{\delta\mu,\gamma_n I_{\mu_0}}$. As they have the same center, we may translate them both to the origin. The error is:

$$\begin{aligned}
\epsilon_{\text{classical}} =& \int_{\mathbb{R}^d}\left| |2\pi I_{\mu_0}|^{-1/2} e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}} - |2\pi\gamma_n I_{\mu_0}|^{-1/2} e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}}\right| d\mathbf{x} && (43)\\
\le& |2\pi I_{\mu_0}|^{-1/2}\int_{\mathbb{R}^d}\left|e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}} - e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}}\right| d\mathbf{x} + 2\pi n^{-\delta/2}\int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma_n}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}} d\mathbf{x} && (44)\\
\le& O(n^{-\delta/2})\int_{\mathbb{R}^d}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x} e^{-\frac{1}{2}\mathbf{x}^T I_{\mu_0}^{-1}\mathbf{x}} d\mathbf{x} && (45)\\
=& O(n^{-\delta/2}) && (46)
\end{aligned}$$

Now we check the quantum term. On the quantum register, the amplifier acts independently on each mode as the displaced thermal state amplifier defined by Eq. (13). From a similar calculation as Eq. (15), we get that

$$\epsilon_{\text{quantum}} \le \frac{1}{2}\sum_{j<k}\left\|\mathcal{A}^{\gamma_n}\left(\rho_{\alpha_{j,k},\beta_{j,k}}\right) - \rho_{\alpha_{j,k},\beta_{j,k}}\right\|_1, \tag{47}$$

where the error of each quantum amplifier is

$$\frac{1}{2}\left\|\mathcal{A}^{\gamma_n}\left(\rho_{\alpha_{j,k},\beta_{j,k}}\right) - \rho_{\alpha_{j,k},\beta_{j,k}}\right\|_1 = O\left(n^{-\delta/2}\right). \tag{48}$$

Therefore, we conclude that the amplification error scales at most as

$$\epsilon_{\text{amp}} \le \epsilon_{\text{classical}} + \epsilon_{\text{quantum}} = O\left(n^{-\delta/2}\right). \tag{49}$$

Let us now consider the term for the Gaussian state compression, which can be expressed as

$$\epsilon_{\text{G}} \le \frac{1}{2}\left\|\mathcal{P}_{\text{c}}\left(N_{\delta\mu,I_{\mu_0}}\right) - N_{\delta\mu,I_{\mu_0}}\right\|_1 + \frac{1}{2}\sum_{j<k}\left\|\mathcal{P}_{n^\delta}\left(\rho_{\alpha_{j,k},\beta_{j,k}}\right) - \rho_{\alpha_{j,k},\beta_{j,k}}\right\|_1 \tag{50}$$

For the classical part, noticing that $\|\delta\mu\|_\infty \le n^{\delta/3}$, we have

$$\begin{aligned}
\left\|\mathcal{P}_{\text{c}}\left(N_{\delta\mu,I_{\mu_0}}\right) - N_{\delta\mu,I_\mu}\right\|_1 \le& \sum_{\|u\|_\infty > n^\delta} N_{\delta\mu,I_{\mu_0}}(u) && (51)\\
\le& \sum_{\|u-\delta\mu\|_\infty > n^\delta - n^{\delta/3}} N_{\delta\mu,I_{\mu_0}}(u) && (52)\\
=& e^{-\Omega(n^{2\delta})} && (53)
\end{aligned}$$

where $N_{\delta\mu,I_{\mu_0}}(u)$ denotes the probability distribution function. For each of the quantum modes, employing Lemma 2, with N substituted by n^δ and $|\alpha_{j,k}| = O(n^{\delta/3})$, we have

$$\frac{1}{2}\left\|\mathcal{P}_{j,k}\left(\rho_{\alpha_{j,k},\beta_{j,k}}\right) - \rho_{\alpha_{j,k},\beta_{j,k}}\right\|_1 = \beta_{j,k}^{\Omega(n^{\delta/8})} + e^{-\Omega(n^{\delta/4})}. \tag{54}$$

Substituting Eqs. (53) and (54) into Eq. (50), we have

$$\epsilon_{\text{G}} = \max_{j<k}\left[\frac{(\mu_0)_k}{(\mu_0)_j}\right]^{\Omega(n^{\delta/8})} + e^{-\Omega(n^{\delta/4})}. \tag{55}$$

Finally, we note that the error of the Q-LAN approximation, corresponding to the errors generated by the transformations between the input state and its Gaussian state approximation, is given by Eqs. (24) and (25) as

$$\epsilon_{\mathrm{Q-LAN}} = O\left(n^{-\kappa(\delta)}\right). \tag{56}$$

Summarizing the above bounds (42), (49), (55), (56) on each of the error terms, we conclude that the protocol generates an error which scales at most

$$\epsilon = O\left(n^{-\kappa(\delta)}\right) + O\left(n^{-\delta/2}\right) \tag{57}$$

C. Memory cost.

There are three sources of memory cost: a classical memory of $[(f_c + f_q)/2]\log n$ bits for the tomography outcome, a classical memory of $f_c\delta\log n$ bits for the classical part of the Gaussian state and a quantum memory of $f_q\delta\log n$ qubits for the quantum part of the Gaussian state. It takes $(1/2+\delta)\log n$ bits to encode a classical free parameter and $(1/2)\log n$ bits plus $\delta\log n$ qubits to encode a quantum free parameter.

From the above discussion we can see that the ratio between the quantum memory cost and the classical memory cost is

$$R_{q/c} = \frac{\delta f_q}{(1/2+\delta)f_c + (1/2)f_q}, \tag{58}$$

which can be made close to zero when δ is set close to zero. This result shows that the quantum memory cost can be made arbitrarily small compared to the classical memory cost.

V. NECESSITY OF A QUANTUM MEMORY.

In the previous section, we showed that the ratio between the quantum and the classical memory cost can be made arbitrarily close to zero [see Eq. (58)]. It is then intuitive to ask whether this ratio can be made equal to zero, i.e. compressing faithfully using a fully classical memory.

The answer to the above question is negative: Here we prove that a protocol using a fully classical memory cannot be faithful, and thus a quantum memory is necessary. To achieve this proof, we first define two distance measures for quantum states as

$$d_H(\rho_1,\rho_2) := \sqrt{2 - 2\,\mathrm{Tr}\left(\rho_1^{1/2}\rho_2^{1/2}\right)} \tag{59}$$

$$d_B(\rho_1,\rho_2) := \sqrt{2 - 2\,\mathrm{Tr}\left|\rho_1^{1/2}\rho_2^{1/2}\right|}. \tag{60}$$

A few properties of these measures will be utilized in our proof: both measures are monotone, namely that they are non-increasing under evolutions of the states [24], [25]. Moreover, $d_H \geq d_B$, and the equality holds if and only if the states are classical.

Lemma 3. *$d_H(\rho_1,\rho_2) \geq d_B(\rho_1,\rho_2)$; the equality holds if and only if $[\rho_1,\rho_2]=0$.*

Proof. First notice that $d_H(\rho_1,\rho_2) \geq d_B(\rho_1,\rho_2)$ if and only if

$$\mathrm{Tr}\left|\rho_1^{1/2}\rho_2^{1/2}\right| \geq \mathrm{Tr}\left(\rho_1^{1/2}\rho_2^{1/2}\right).$$

Consider the singular value decomposition $\rho_1^{1/2}\rho_2^{1/2} = \sum_i s_i|\psi_i\rangle\langle\phi_i|$. We have $\mathrm{Tr}\left|\rho_1^{1/2}\rho_2^{1/2}\right| = \sum_i s_i$ and $\mathrm{Tr}\left(\rho_1^{1/2}\rho_2^{1/2}\right) = \sum_i s_i\langle\phi_i|\psi_i\rangle$. The former is always an upper bound of the latter, which proves $d_H \geq d_B$. Moreover, these two quantities are equal if and only if $\langle\psi_i|\phi_i\rangle = 1$ for every i, namely that $\rho_1^{1/2}\rho_2^{1/2}$ is positive definite. Finally, we have

$$\left(\rho_1^{1/2}\rho_2^{1/2}\right)^\dagger = \left(\rho_1^{1/2}\rho_2^{1/2}\right) \Leftrightarrow \left[\rho_1^{1/2},\rho_2^{1/2}\right] = 0,$$

which further implies that $[\rho_1,\rho_2]=0$. □

We now prove that fully classical memory cannot fulfill the task of faithful compression. When a compression protocol $(\mathcal{E},\mathcal{D})$ uses only classical memory, the encoder $\mathcal{E}$ is essentially a quantum-classical channel, turning quantum states into classical distributions, while the decoder $\mathcal{D}$ is a classical-quantum channel, turning classical distributions back into quantum states. We prove the following lemma, which states that d_H is approximately equal to d_B for states that can be faithfully compressed using classical memory.

Lemma 4. *If $\mathcal{E}$ is a quantum-classical channel and $\mathcal{D}$ is a classical-quantum channel satisfying $\|\mathcal{D}\circ\mathcal{E}(\rho_i) - \rho_i\|_1 \leq \epsilon$ for $i=1,2$, the following inequality holds*

$$d_H(\rho_1,\rho_2) \leq d_B(\rho_1,\rho_2) + 2\sqrt{\epsilon}. \tag{61}$$

Proof. First notice that, since $\mathcal{E}$ is a quantum-classical channel, we have $[\mathcal{E}(\rho_1),\mathcal{E}(\rho_2)] = 0$. Applying Lemma 3, we immediately get that

$$d_H\left(\mathcal{E}(\rho_1),\mathcal{E}(\rho_2)\right) = d_B\left(\mathcal{E}(\rho_1),\mathcal{E}(\rho_2)\right). \tag{62}$$

Using the monotonicity of d_H and d_B, we have

$$d_H\left(\mathcal{D}\circ\mathcal{E}(\rho_1),\mathcal{D}\circ\mathcal{E}(\rho_2)\right) \le d_B\left(\rho_1,\rho_2\right). \tag{63}$$

Substituting the triangle inequality

$$d_H(\rho_1,\rho_2) \le d_H\left(\mathcal{D}\circ\mathcal{E}(\rho_1),\mathcal{D}\circ\mathcal{E}(\rho_2)\right) + d_H\left(\rho_1,\mathcal{D}\circ\mathcal{E}(\rho_1)\right) + d_H\left(\mathcal{D}\circ\mathcal{E}(\rho_2),\rho_2\right)$$

into the above inequality, we have

$$d_H(\rho_1,\rho_2) \le d_B\left(\rho_1,\rho_2\right) + d_H\left(\rho_1,\mathcal{D}\circ\mathcal{E}(\rho_1)\right) + d_H\left(\mathcal{D}\circ\mathcal{E}(\rho_2),\rho_2\right). \tag{64}$$

Since d_H is upper bounded by the trace distance as $d_H(\rho,\sigma) \le \sqrt{\|\rho-\sigma\|_1}$ [26], we have

$$d_H(\rho_1,\rho_2) \le d_B\left(\rho_1,\rho_2\right) + 2\sqrt{\epsilon} \tag{65}$$

as desired.

□

Now, suppose a qudit state subfamily $\{\rho_\theta^{\otimes n}\}$, with θ containing a free quantum parameter for unitary rotation, can be faithfully compressed using only classical memory. We can use the above lemmas to show a contradiction.

Consider two states $\rho_\theta^{\otimes n}$ and $\rho_{\theta+s/\sqrt{n}}^{\otimes n}$ from the family, where all entries of s are zero except for the one for the free quantum parameter. Applying Q-LAN to a neighborhood of θ, we know that these two states can be converted into a thermal state $\rho_\beta^{(\mathrm{thm})}$ and $\rho_{\alpha,\beta}$ reversibly (up to a vanishing error), with $\beta=\beta(\theta)$ and $\alpha=\alpha(\theta,s)$ independent of n. This implies that there exists a quantum-classical channel and a classical-quantum channel satisfying the conditions in Lemma 4, with $\rho_1=\rho_\beta^{(\mathrm{thm})}$, $\rho_2=\rho_{\alpha,\beta}$ and ϵ vanishing in n, namely

$$d_H\left(\rho_\beta^{(\mathrm{thm})},\rho_{\alpha,\beta}\right) \le d_B\left(\rho_\beta^{(\mathrm{thm})},\rho_{\alpha,\beta}\right) + 2\sqrt{\epsilon_n} \tag{66}$$

with $\lim_{n\to\infty}\epsilon_n = 0$. Taking the limit $n\to\infty$ on both sides of the inequality and using Lemma 3, we know that $d_B = d_H$ for these two states, which implies that

$$\left[\rho_\beta^{(\mathrm{thm})},\rho_{\alpha,\beta}\right] = 0. \tag{67}$$

It is easy to verify that the thermal state $\rho_\beta^{(\mathrm{thm})}$ and its displaced counterpart $\rho_{\alpha,\beta} = D_\alpha\rho_\beta^{(\mathrm{thm})}D_\alpha^\dagger$ do not commute:

$$\langle m|\left[\rho_\beta^{(\mathrm{thm})},\rho_{\alpha,\beta}\right]|l\rangle = (1-\beta)(\beta^m-\beta^l)\langle m|\rho_{\alpha,\beta}|l\rangle$$

which is not zero when $m\neq l$, and thus we have a contradiction. At last, we note that it can proved in the same way that quantum memory is necessary in the displaced thermal state compression.

VI. Optimality of the compression

In this section, we prove the optimality of our compression protocol, namely that any protocol with an overall memory size of $(f/2-\delta)\log n$ for $\delta>0$ cannot be faithful for a subfamily of f free parameters. The idea is to construct a communication protocol which conveys nearly $(f/2)\log n$ bits, using the compression protocol. From the well-known principle by Holevo [27], we know that such an amount of bits cannot be encoded in less than $(f/2)\log n$ qubits, and thus the memory size of the protocol is at least $(f/2)\log n$ even if the memory is fully quantum.

We begin by defining a mesh M on the parameter space Θ, which is a collection of points corresponding to states that are almost mutually distinguishable:

$$\mathsf{M} = \left\{\theta\in\Theta \mid |(\theta-\theta_0)_i| = z_i\cdot\log n/\sqrt{n}, z_i\in\mathbb{N}\ \forall\ i\right\} \tag{68}$$

where $\theta_0\in\Theta$ is a fixed point. The mesh M is so defined that

$$\epsilon_{\min} := \min_{\theta\neq\theta'\in\mathsf{M}}\frac{1}{2}\|\rho_\theta-\rho_{\theta'}\|_1 = \frac{C\log n}{2\sqrt{n}} + O\left(\frac{\log^2 n}{n}\right), \tag{69}$$

which comes from the Euclidean expansion of trace distance, namely that $\|\rho_\theta-\rho_{\theta'}\|_1 = C\|\theta-\theta'\|_\infty + O(\|\theta-\theta'\|_\infty^2)$ for a constant $C>0$. On the other hand, the number of points contained in the mesh satisfies

$$|\mathsf{M}| \ge T_\Theta\left[\frac{\sqrt{n}}{\log n}\right]^f, \tag{70}$$

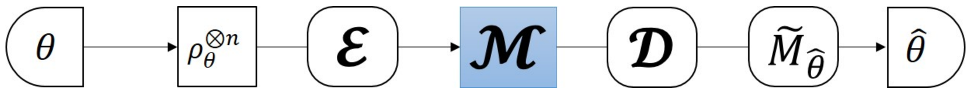

Fig. 4. **A protocol to communicate** $\log|\mathcal{M}|$ **bits of information.** Here $\mathcal{E}$ and $\mathcal{D}$ are the encoder and the decoder, $\tilde{M}_{\hat{\theta}}$ is the POVM to recover the message, and $\mathcal{M}$ denotes the memory.

where $T_\Theta > 0$ is independent of n.

Now, if a compression protocol $(\mathcal{E}, \mathcal{D})$ with a memory size $n_{\rm enc}$ is faithful, two parties can apply the following protocol to communicate $\log|\mathsf{M}|$ bits of classical information almost perfectly (see also Fig. 4):

1) Both parties agree apriori on a one-to-one correspondence between messages and point in the mesh M. To communicate a certain message, the sender picks the corresponding point θ on the mesh and prepares $\rho_\theta^{\otimes n}$.
2) The sender applies the encoder $\mathcal{E}$ and transmits $\mathcal{E}(\rho_\theta^{\otimes n})$ to the receiver.
3) The receiver recovers the message by first applying the decoder $\mathcal{D}$ and then measuring with the POVM $\{\tilde{M}_{\hat{\theta}}\}$ defined as
$$\tilde{M}_{\hat{\theta}} = \int_{\frac{1}{2}\|\sigma-\rho_{\hat{\theta}}\|_1 \le \frac{\epsilon_{\min}}{2}} \mathrm{d}\,\sigma\, M_\sigma \tag{71}$$
where M_σ is the qudit tomography POVM element defined in Eq. (9) of [23], which the property that
$$\int_{\frac{1}{2}\|\sigma-\rho\|_1 \le \epsilon/2} \mathrm{d}\,\sigma\ \mathrm{Tr}[M_\sigma\, \rho^{\otimes n}] \ge 1-(n+1)^{3d^2} e^{-n\epsilon^2}. \tag{72}$$
(For displaced thermal state families, a similar result holds for the heterodyne measurement of α and maximum likelihood estimation of β [19].)

The point-wise probability of error for the above protocol can be bounded as
$$\begin{aligned} P_e(\theta) &= \int_{\frac{1}{2}\|\rho_{\theta'}-\sigma\|_1 > \frac{\epsilon_{\min}}{2}} \mathrm{d}\,\sigma\ \mathrm{Tr}\left[\rho_\theta^{\otimes n}\, M_\sigma\right] && (73)\\ &\le (n+1)^{3d^2} e^{-\frac{C\log^2 n}{4}} && (74)\\ &\le n^{-\frac{C\log n}{8}}. && (75)\end{aligned}$$

Consider the case when the messages are uniformly distributed. Using the monotonicity of mutual information and the upper bound of entropy, we have
$$\begin{aligned} n_{\rm enc} &\ge H\left[\mathcal{E}\left(|\mathsf{M}|^{-1}\sum_{\theta\in\mathsf{M}} \rho_\theta^{\otimes n}\right)\right]\\ &\ge I(\Theta:\mathcal{M})\\ &\ge I(\Theta:\hat{\Theta}). \end{aligned} \tag{76}$$

Applying Fano's inequality [28], the bound continues as
$$\begin{aligned} n_{\rm enc} &\ge (1-P_e)\log|\mathsf{M}| - h(P_e)\\ &= \frac{f}{2}\log n - f\log\log n + o(1), \end{aligned} \tag{77}$$
where ϵ is the error of compression, $h(x) = -x\log x$ $(h(0) := 0)$ and $P_e \le n^{-\frac{C\log n}{8}}$ is the average error probability. The optimality of our compression protocol is justified by Eq. (77).

VII. Conclusion

In this work we have solved the problem of compressing identically prepared states of finite-dimensional quantum systems and identically prepared displaced thermal states. We showed that the total size of the required memory is proportional to the number of free parameters of the state. Moreover, we observed the asymptotic ratio between the amount of quantum bits and the amount of classical bits can be set to an arbitrarily small constant. Still, a fully classical memory cannot faithfully encode genuine quantum states: only states that are jointly diagonal in fixed basis can be compressed into a purely classical memory.

A natural development of our work is the study of compression protocols for quantum population coding beyond the case of identically prepared states. Motivated by the existing literature on classical population coding [9], the idea is to consider families of states representing a population of quantum particles. At the most fundamental level, the indistinguishability of quantum particles leads to the Bose-Einstein [29], [30] and Fermi-Dirac statistics [31], [32], as well as to other intermediate statistics [33], [34]. Since the Hilbert space describing identical quantum particles is not the tensor product of single-particle Hilbert spaces, the compression of quantum populations of indistinguishable particles requires a non-trivial extension of our results. The optimal compression protocols are likely to shed light on how information is encoded into a broad range of real physical systems. In addition, the compression protocols will offer a tool to simulate large numbers of particles using quantum computers of relatively smaller size.

From the point of view of quantum simulations, it is also meaningful to consider the compression of tensor network states [35], [36] which provide a variational ansatz for a large number of quantum manybody systems. The extension of quantum compression to this scenario is appealing as a technique to reduce the number of qubits needed to simulate systems of distinguishable quantum particles, in the same spirit of the compressed simulations introduced by Kraus [37] and coauthors [38], [39] for the Ising model and other models in quantum statistical mechanics. In the long term, the information-theoretic study of manybody quantum systems may provide a new approach to the simulation of complex systems that are not efficiently simulatable on classical computers.

Acknowledgments

This work is supported by the Canadian Institute for Advanced Research (CIFAR), by the Hong Kong Research Grant Council through Grant No. 17326616, by National Science Foundation of China through Grant No. 11675136, and by the HKU Seed Funding for Basic Research, and the John Templeton Foundation. Y. Y. is supported by a Hong Kong and China Gas Scholarship. MH was supported in part by a MEXT Grant-in-Aid for Scientific Research (B) No. 16KT0017, the Okawa Research Grant and Kayamori Foundation of Informational Science Advancement.

References

[1] Konrad Banaszek, Marcus Cramer, and David Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.
[2] Valerio Scarani, Sofyan Iblisdir, Nicolas Gisin, and Antonio Acín. Quantum cloning. *Reviews of Modern Physics*, 77:1225–1256, Nov 2005.
[3] Nicolas J Cerf, A Ipe, and Xavier Rottenberg. Cloning of continuous quantum variables. *Physical Review Letters*, 85(8):1754, 2000.
[4] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009.
[5] Yuxiang Yang, Giulio Chiribella, and Gerardo Adesso. Certifying quantumness: Benchmarks for the optimal processing of generalized coherent and squeezed states. *Physical Review A*, 90(4):042319, 2014.
[6] Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(4):2738, 1995.
[7] Richard Jozsa and Benjamin Schumacher. A new proof of the quantum noiseless coding theorem. *Journal of Modern Optics*, 41(12):2343–2349, 1994.
[8] Hoi-Kwong Lo. Quantum coding theorem for mixed states. *Optics Communications*, 119(5-6):552–556, 1995.
[9] Si Wu, Shun-ichi Amari, and Hiroyuki Nakahara. Population coding and decoding in a neural field: a computational study. *Neural Computation*, 14(5):999–1026, 2002.
[10] Martin Plesch and Vladimír Bužek. Efficient compression of quantum information. *Physical Review A*, 81(3):032317, 2010.
[11] Lee A. Rozema, Dylan H. Mahler, Alex Hayat, Peter S. Turner, and Aephraim M. Steinberg. Quantum data compression of a qubit ensemble. *Physical Review Letters*, 113:160504, Oct 2014.
[12] Yuxiang Yang, Giulio Chiribella, and Daniel Ebler. Efficient quantum compression for ensembles of identically prepared mixed states. *Physical Review Letters*, 116:080501, Feb 2016.
[13] Yuxiang Yang, Giulio Chiribella, and Masahito Hayashi. Optimal compression for identically prepared qubit states. *Physical Review Letters*, 117:090502, Aug 2016.
[14] Masahito Hayashi and Vincent Tan. Minimum rates of approximate sufficient statistics. *arXiv preprint arXiv: 1612.02542*, 2016.
[15] Masahito Hayashi and Keiji Matsumoto. Asymptotic performance of optimal state estimation in qubit system. *Journal of Mathematical Physics*, 49(10):102101, 2008.
[16] Mădălin Guţă and Jonas Kahn. Local asymptotic normality for qubit states. *Physical Review A*, 73(5):052108, 2006.
[17] Mădălin Guţă and Anna Jenčová. Local asymptotic normality in quantum statistics. *Communications in Mathematical Physics*, 276(2):341–379, 2007.
[18] Jonas Kahn and Mădălin Guţă. Local asymptotic normality for finite dimensional quantum systems. *Communications in Mathematical Physics*, 289(2):597–652, 2009.
[19] Masahito Hayashi. Asymptotic quantum estimation theory for the thermal states family. In *Quantum Communication, Computing, and Measurement 2*, pages 99–104. Springer, 2002.
[20] Wataru Kumagai and Masahito Hayashi. Quantum hypothesis testing for gaussian states: Quantum analogues of χ^2, t-, and f-tests. *Communications in Mathematical Physics*, 318(2):535–574, 2013.
[21] Carlton M Caves. Quantum limits on noise in linear amplifiers. *Physical Review D*, 26(8):1817, 1982.
[22] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
[23] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 913–925, New York, NY, USA, 2016. ACM.
[24] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
[25] Shunlong Luo and Qiang Zhang. Informational distance on quantum-state space. *Physical Review A*, 69(3):032106, 2004.
[26] Alexander Semenovich Holevo. On quasiequivalence of locally normal states. *Theoretical and Mathematical Physics*, 13(2):1071–1082, 1972.
[27] Alexander Semenovich Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.
[28] Robert M Fano and David Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.
[29] Satyendra Nath Bose. Plancks gesetz und lichtquantenhypothese. *Z. phys*, 26(3):178, 1924.
[30] Albert Einstein. *Quantentheorie des einatomigen idealen Gases*. Akademie der Wissenshaften, in Kommission bei W. de Gruyter, 1924.
[31] Enrico Fermi. Eine statistische methode zur bestimmung einiger eigenschaften des atoms und ihre anwendung auf die theorie des periodischen systems der elemente. *Zeitschrift für Physik*, 48(1-2):73–79, 1928.
[32] Paul AM Dirac. The quantum theory of the electron. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 117, pages 610–624. The Royal Society, 1928.
[33] Jon M Leinaas and Jan Myrheim. On the theory of identical particles. *Il Nuovo Cimento B (1971-1996)*, 37(1):1–23, 1977.
[34] Frank Wilczek. Quantum mechanics of fractional-spin particles. *Physical review letters*, 49(14):957, 1982.

[35] Norbert Schuch, David Pérez-García, and Ignacio Cirac. Classifying quantum phases using matrix product states and projected entangled pair states. *Physical Review B*, 84(16):165139, 2011.
[36] Norbert Schuch, Michael M Wolf, Frank Verstraete, and J Ignacio Cirac. Entropy scaling and simulability by matrix product states. *Physical Review Letters*, 100(3):030504, 2008.
[37] Barbara Kraus. Compressed quantum simulation of the ising model. *Physical review letters*, 107(25):250503, 2011.
[38] Walter León Boyajian, Valentin Murg, and Barbara Kraus. Compressed simulation of evolutions of the x y model. *Physical Review A*, 88(5):052329, 2013.
[39] Walter León Boyajian and Barbara Kraus. Compressed simulation of thermal and excited states of the one-dimensional x y model. *Physical Review A*, 92(3):032323, 2015.
[40] Andreas Winter. Coding theorem and strong converse for quantum channels. *IEEE Transactions on information theory*, 45(7):2481, 1999.
[41] Mark Wilde. The quantum data-compression theorem. In *Quantum information theory*, chapter 18. Cambridge University Press, 2013.
[42] FAM De Oliveira, Myungshik Kim, Peter L Knight, and Vladimir Bužek. Properties of displaced number states. *Physical Review A*, 41(5):2645, 1990.

Appendix

A. Proof of Lemma 1.

For any $x>0$, we define a series of $t+1=\lfloor n^{1/2+x}\rfloor$ intervals $\mathsf{L}_0,\dots \mathsf{L}_t$ as

$$\begin{aligned}
&\mathsf{L}_0=\{0\}\\
&\mathsf{L}_i=\{(i-1)\lfloor n^{(1-x)/2}\rfloor+1,\dots,i\lfloor n^{(1-x)/2}\rfloor\}\quad 0<i<t\\
&\mathsf{L}_t=\{(t-1)\lfloor n^{(1-x)/2}\rfloor+1,\dots\}.
\end{aligned}$$

For any non-negative integer m, we denote by $i(m)$ the index of the interval containing m, i.e. $m\in\mathsf{L}_{i(m)}$.

To design the compression protocol, we first notice that the n-fold thermal state can be written in the form

$$\left(\rho_\beta^{(\mathrm{thm})}\right)^{\otimes n}=(1-\beta)^n\sum_{\vec m}\beta^{|\vec m|}|\vec m\rangle\langle\vec m|,\tag{78}$$

where $|\vec m\rangle=|m_1\rangle\otimes\cdots\otimes|m_n\rangle$ is the photon number basis of n modes and $|\vec m|:=m_1+\cdots+m_n$. The compression protocol runs as follows:

- *Encoder*. First perform projective measurement in the photon number basis of n modes, which yields an n-dimensional vector $\vec m$. Then compute $i(|\vec m|)$ and encode it into a classical memory. The encoding channel can be represented as

$$\mathcal{E}_{n,x}^{(\mathrm{thm})}(\rho):=\sum_{\vec m}\langle\vec m|\rho|\vec m\rangle\ |i(|\vec m|)\rangle\langle i(|\vec m|)|.$$

- *Decoder*. Read an integer i from the memory. If $i\ge t$ prepare a fixed state $|\vec t\rangle\langle\vec t|$ (defined below); if $i<t$ perform random sampling in the interval L_i. For each outcome $\hat m$ of the sampling, prepare the n-mode state

$$\binom{n+\hat m-1}{\hat m}^{-1}\sum_{\vec m:|\vec m|=\hat m}|\vec m\rangle\langle\vec m|.$$

  Then the decoding channel can be represented as

$$\mathcal{D}_{n,x}^{(\mathrm{thm})}\left(|i\rangle\langle i|\right):=\begin{cases}\sum_{\vec m:|\vec m|\in\mathsf{L}_i}\frac{|\vec m\rangle\langle\vec m|}{|\mathsf{L}_i|\binom{n+|\vec m|-1}{|\vec m|}} & i<t\\ |\vec t\rangle\langle\vec t| & i=t\end{cases}$$
$$\vec t=((t-1)\lfloor n^{(1-x)/2}\rfloor+1,\dots,(t-1)\lfloor n^{(1-x)/2}\rfloor+1).$$

It is straightforward from definition that the protocol uses $\log(t+1)=(1/2+x)\log n+o(1)$ classical bits. What remains is to bound the error of the protocol. First, we notice that the recovered state is

$$\mathcal{D}_{n,x}^{(\mathrm{thm})}\circ\mathcal{E}_{n,x}^{(\mathrm{thm})}\left[\left(\rho_\beta^{(\mathrm{thm})}\right)^{\otimes n}\right]=\sum_{\vec m}\left(\rho_{\beta,n}\right)_{\vec m}|\vec m\rangle\langle\vec m|\tag{79}$$

$$\left(\rho_{\beta,n}\right)_{\vec m}=\begin{cases}(1-\beta)^n\beta^{|\vec m|}\sum_{m\in\mathsf{L}_{i(|\vec m|)}}\frac{\beta^{m-|\vec m|}\binom{n+m-1}{m}}{|\mathsf{L}_{i(|\vec m|)}|\binom{n+|\vec m|-1}{|\vec m|}} & i(|\vec m|)<t\\[2ex] \sum_{\vec m':i(|\vec m'|)=t}(1-\beta)^n\beta^{|\vec m'|} & \vec m=\vec t\\[2ex] 0 & \text{else.}\end{cases}\tag{80}$$

We choose S as the minimal set satisfying i) S is a union of several intervals chosen from the set $\{\mathsf{L}_i\}$ and ii)

$$\mathsf{S}\supset\left[\frac{\beta n}{1-\beta}-n^{(1-x)/2},\frac{\beta n}{1-\beta}+n^{(1-x)/2}\right].\tag{81}$$

Apparently, $\mathsf{S} \subset [0, n^{1+x/2}]$ for large enough n. Then the error can be bounded as

$$\begin{aligned}
\epsilon_{\mathrm{thm}} &= \frac{1}{2}\left\|\mathcal{D}^{(\mathrm{thm})}_{n,x}\circ\mathcal{E}^{(\mathrm{thm})}_{n,x}\left[\left(\rho^{(\mathrm{thm})}_{\beta}\right)^{\otimes n}\right]-\left(\rho^{(\mathrm{thm})}_{\beta}\right)^{\otimes n}\right\|_1\\
&= \frac{1}{2}\sum_{\vec{m}}\left|(1-\beta)^n\beta^{|\vec{m}|}-(\rho_{\beta,n})_{\vec{m}}\right|\\
&\le \sum_{\vec{m}:|\vec{m}|\notin\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}+\frac{1}{2}\left[\sum_{\vec{m}:|\vec{m}|\in\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}\right]\max_{\substack{m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\sum_{m\in\mathsf{L}_i}\frac{\beta^m\binom{n+m-1}{m}}{|\mathsf{L}_i|\beta^{m'}\binom{n+m'-1}{m'}}-1\right|+\sum_{\vec{m}':i(|\vec{m}'|)=t}(1-\beta)^n\beta^{|\vec{m}'|}\\
&\le \sum_{\vec{m}:|\vec{m}|\notin\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}+\frac{1}{2}\left[\sum_{\vec{m}:|\vec{m}|\in\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}\right]\max_{\substack{m,\,m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\frac{\beta^m\binom{n+m-1}{m}}{\beta^{m'}\binom{n+m'-1}{m'}}-1\right|+O\left((1-\beta)^n\beta^n\right)\\
&\le \sum_{\vec{m}:|\vec{m}|\notin\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}+\frac{1}{2}\max_{\substack{m,\,m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\frac{\beta^m\binom{n+m-1}{m}}{\beta^{m'}\binom{n+m'-1}{m'}}-1\right|+O\left((1-\beta)^n\beta^n\right).
\end{aligned}$$

On one hand, we notice that $|\vec{m}|$ is the sum of n i.i.d. random variables with geometric distribution $\{(1-\beta)\beta^i\}_{i=0}^{\infty}$ and thus, by Central Limit Theorem, the first error term scales as

$$\sum_{\vec{m}:|\vec{m}|\notin\mathsf{S}}(1-\beta)^n\beta^{|\vec{m}|}=O\left[\mathrm{erfc}\left(\frac{n^{x/2}(1-\beta)}{\sqrt{2\beta}}\right)\right]=e^{-\Omega(n^x)}$$

where $\mathrm{erfc}(x):=(2/\pi)\int_x^{\infty}e^{-s^2}\,\mathrm{d}\,s$ is the complementary error function. On the other hand, in the second error term, m and m' are in the same order as n, so the second term can be bounded as

$$\begin{aligned}
\max_{\substack{m,\,m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\frac{\beta^m\binom{n+m-1}{m}}{\beta^{m'}\binom{n+m'-1}{m'}}-1\right| &= \max_{\substack{m,\,m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\beta^{m-m'}\frac{(n+m-1)\cdots(n+m')}{m\ldots(m'+1)}-1\right|\\
&= \max_{\substack{m,\,m'\in\mathsf{L}_i\\ \mathsf{L}_i\cap\mathsf{S}\neq\emptyset}}\left|\left(\frac{\beta m+\beta n}{m}\right)^{m-m'}\left[1+O\left(\frac{|\mathsf{L}_i|^2}{n}\right)\right]-1\right|\\
&=\left|\left[1+O\left(n^{-x}\right)\right]\left[1+O\left(n^{-x}\right)\right]-1\right|\\
&=O\left(n^{-x}\right).
\end{aligned}$$

Therefore, we have proved Eq. (9).

B. Proof of Lemma 2.

For any input state ρ we have

$$\begin{aligned}
\epsilon(\rho) &\le \frac{1}{2}\left\|P_N\rho P_N-\rho\right\|_1+\frac{1}{2}[1-\mathrm{Tr}(\rho P_N)] && (82)\\
&\le \frac{1}{2}\left[2\sqrt{[1-\mathrm{Tr}(\rho P_N)]}+1-\mathrm{Tr}(\rho P_N)\right] && (83)\\
&\le \frac{3}{2}\sqrt{1-\mathrm{Tr}(\rho P_N)}. && (84)
\end{aligned}$$

The second inequality came from the gentle measurement lemma [40], [41]. Substituting $\rho_{\alpha,\beta}=D_\alpha\rho^{(\mathrm{thm})}_{\beta}D_\alpha^{\dagger}$ into the above bound, we express the truncation error for $\rho_{\alpha,\beta}$ as

$$\epsilon(\rho_{\alpha,\beta})\le\frac{3}{2}\sqrt{1-\mathrm{Tr}\left[D_\alpha\rho^{(\mathrm{thm})}_{\beta}D_\alpha^{\dagger}P_N\right]}. \tag{85}$$

We now bound the trace part in the right hand side of the last inequality as

$$\begin{aligned}
1-\mathrm{Tr}\left[D_\alpha\rho^{(\mathrm{thm})}_{\beta}D_\alpha^{\dagger}P_\alpha\right] &= \sum_{k\le N}\sum_{l=0}^{\infty}(1-\beta)\beta^l|\langle l|D_\alpha|k\rangle|^2\\
&\le \max_{l\le l_0}\sum_{k>N}|\langle l|D_\alpha|k\rangle|^2+\sum_{l\ge l_0}^{\infty}(1-\beta)\beta^l\\
&\le \max_{l\le l_0}\sum_{k>N}|\langle l|D_\alpha|k\rangle|^2+\beta^{l_0} && (86)
\end{aligned}$$

Here we set $l_0 = N^{x/8}$. Notice that $|\langle l|D_\alpha|k\rangle|^2$ is the photon number distribution of a displaced number state [42], which can be bounded as

$$\begin{aligned}|\langle l|D_\alpha|k\rangle|^2 =& \frac{e^{-|\alpha|^2}(|\alpha|^2)^{k+l}}{k!l!}\left|\sum_{i=0}^{\min\{k,l\}}\frac{k!l!(-|\alpha|^2)^{-i}}{i!(k-i)!(l-i)!}\right|^2\\ \leq& \frac{e^{-|\alpha|^2}(|\alpha|^2)^{k+l}}{k!l!}\left|\sum_{i=0}^{k}\binom{k}{i}\left(\frac{l}{|\alpha|^2}\right)^i\right|^2\\ =& \frac{e^{-|\alpha|^2}|\alpha|^{2l}}{k!l!}\left(\frac{l+|\alpha|^2}{|\alpha|}\right)^{2k}.\end{aligned}$$

Then we can bound the first term in (86) as

$$\begin{aligned}\max_{l\leq l_0}\sum_{k>N}|\langle l|D_\alpha|k\rangle|^2 \leq& \max_{l\leq l_0}\frac{|\alpha|^{2l}}{l!}\sum_{k>N}\frac{e^{-|\alpha|^2}}{k!}\left(\frac{l+|\alpha|^2}{|\alpha|}\right)^{2k}\\ =& \max_{l\leq l_0}\frac{e^{2l+\frac{l^2}{|\alpha|^2}}|\alpha|^{2l}}{l!}\sum_{k>N}\mathbf{Pois}_{\lambda_\alpha}(k),\end{aligned}$$

where $\mathbf{Pois}_{\lambda_\alpha}(k)$ is the Poisson distribution with mean $\lambda_\alpha = (l+|\alpha|^2)^2/|\alpha|^2$. Notice that $[\lambda_\alpha - N^{1/2+x/2}, \lambda_\alpha + N^{1/2+x/2}] \subseteq [0, N]$ and thus we have

$$\begin{aligned}\max_{l\leq l_0}\sum_{k>N}|\langle l|D_\alpha|k\rangle|^2 \leq& \max_{l\leq l_0}\frac{|\alpha|^{2l}e^{2l+\frac{l^2}{|\alpha|^2}}}{l!}\sum_{|k-\lambda_\alpha|>N^{1/2+x/2}}\mathbf{Pois}_{\lambda_\alpha}(k)\\ =& \max_{l\leq|\alpha|^{x/4}}\frac{|\alpha|^{2l}e^{2l+\frac{l^2}{|\alpha|^2}}}{l!}e^{-\Omega(N^{x/2})}\\ =& e^{-\Omega(N^{x/4})}.\end{aligned}$$

having used the tail bound for Poisson distribution and $l_0 = N^{x/8}$. Substituting the above bound into Eq. (86), we get

$$\sqrt{1-\mathrm{Tr}\left[D_\alpha\rho_\beta^{(\mathrm{thm})}D_\alpha^\dagger P_\alpha\right]} \leq \beta^{\Omega(N^{x/8})} + e^{-\Omega(N^{x/4})}.$$

Finally, substituting the above inequality into Eq. (85), we can bound the error as

$$\epsilon(\rho_{\alpha,\beta}) = \beta^{\Omega(N^{x/8})} + e^{-\Omega(N^{x/4})}. \tag{87}$$