Lossless quantum prefix compression for communication channels that are always open

Markus Müller[*]
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

Caroline Rogers[†] and Rajagopal Nagarajan[‡]
Department of Computer Science, University of Warwick Coventry, CV47AL, United Kingdom

(Received 6 September 2008; published 6 January 2009)

We describe a method for lossless quantum compression if the output of the information source is not known. We compute the best possible compression rate, minimizing the expected base length of the output quantum bit string (the base length of a quantum string is the maximal length in the superposition). This complements work by Schumacher and Westmoreland who calculated the corresponding rate for minimizing the output's average length. Our compressed code words are prefix-free indeterminate-length quantum bit strings which can be concatenated in the case of multiple sources. Therefore, we generalize the known theory of prefix-free quantum codes to the case where strings have indeterminate length. Moreover, we describe a communication model which allows the lossless transmission of the compressed code words. The benefit of compression is then the reduction of transmission errors in the presence of noise.

DOI: 10.1103/PhysRevA.79.012302 PACS number(s): 03.67.Hk

I. INTRODUCTION

One of the main aims of information theory is to determine the most efficient way to compress messages. The solution to this problem often reveals relations to entropylike quantities, as in Shannon's noiseless coding theorem [1], where the entropy of the information source determines the best possible compression rate.

The situation in quantum information theory is quite similar. The most popular example is Schumacher's noiseless coding theorem [2], showing that the best possible compression rate in the quantum case is given by von Neumann entropy. "Compression" here means that the number of qubits that have to be transmitted to faithfully exchange a quantum state is minimized. This definition shows that the compression of quantum information is automatically related to the problem of communication: once the compression is accomplished, then *how can the compressed code words be transmitted to a receiver?*

This question addresses an important difficulty in the quantum situation which does not arise in classical information theory: if a variable-length code is used for quantum compression, some code words will be shorter than others. But this may result in code words which are in a superposition of different lengths—how can those code words be transmitted without disturbance?

This problem is one of several reasons why it was previously stated [3–7] that lossless compression of an ensemble $\mathcal{E}=\{p_i,|\psi_i\rangle\}$ of quantum states is in general *impossible* if the value i of the state $|\psi_i\rangle$ to be compressed is unknown. A related objection [7] is that *prefix-free codes* are also useless in the quantum situation: a prefix-free code word carries its own length information. If it is transmitted over a channel, that length information must be read out to see when the transmission is over and the channel can be closed. Again, if the code word is in a superposition of different lengths, this reading-out measurement disturbs the code word.

In this paper, we show that the aforementioned problems do not appear if one uses a channel instead which is *always open*. In this case, there is no need to decide when the transmission is finished. Even in the case of such a channel, compression can be beneficial: it can help to reduce transmission errors.

To better understand the purpose of this paper, it makes sense to think about the compression of quantum information as taking place in several steps.

Step 1. First, the quantum state is *compressed*, typically yielding an output code which is in a superposition of different lengths.

Step 2. Optional: the output code is *cut off* (projected) to get a determinate-length code (which introduces some loss).

Step 3. Finally, the code is *transmitted* over some quantum channel.

Actually, Schumacher and Westmoreland [7] give a method of this form for compression of quantum information using a prefix-free quantum code. In fact, step 1 in their setting is *lossless*—it is a unitary and thus reversible operation that minimizes the output's average length.

Then they show that a projection to the first $n(S+\delta)$ qubits does not disturb the message very much, where S is the source's entropy—this is step 2 in the scheme above—which introduces some (small) loss. As the resulting code word consists of a classically known, determinate number of qubits, it is clear how to transmit it over a channel (step 3).

In this paper, we describe how step 1 can be carried out losslessly if the task is to minimize the output's *base length* instead of the average length (both length notions will be discussed in detail below; cf. Definition III 2), and we compute the best possible compression rate in terms of an entropylike quantity (Theorem VI 4) for the case of a single quantum information source. We do this using *prefix-free*

[*] Also at Institute of Mathematics, Berlin Institute of Technology (TU Berlin). mueller@math.tu-berlin.de

[†] caroline@dcs.warwick.ac.uk

[‡] biju@dcs.warwick.ac.uk

quantum bit strings such that code words can be concatenated in the case of several sources.

For this reason, we advance the theory of prefix-free quantum strings by generalizing the definition and results of Schumacher and Westmoreland in a natural way.

Moreover, we explain that step 3 is unproblematic if the channel in question is *always open*, even if step 2 is dropped. All in all, this gives a *lossless* method of compression and transmission of quantum information. The price we pay for it is that there is no way to see when the transmission is finished. Yet the benefit is that the probability of transmission errors can be reduced.

II. SYNOPSIS

This paper is organized as follows.

In Sec. III, we give a brief description of previous work on lossless quantum compression. In particular, we review the arguments why and in what way lossless quantum compression of unknown states seems to be impossible. We define indeterminate-length quantum bit strings (as used by several authors before) and give a physical interpretation.

In Sec. IV, we give a communication model which describes a situation where lossless quantum compression is possible and useful. In short, we explain the model of an "always-open channel" where neither Alice nor Bob knows when the transmission has finished, but both parties benefit from compression by reducing transmission errors.

Section V contains a review of some results on prefix-free quantum bit strings, generalizing work by Schumacher and Westmoreland [7]. We also give results which have useful interpretations in the framework of our compression scheme. Moreover, we prove that the concatenation of prefix-free indeterminate-length quantum bit strings can in principle be implemented physically.

Our main result is Theorem VI 4. It states the optimal rate for prefix-free compression of the unknown output of a single quantum information source, given the task to minimize the expected base length.

To state the theorem, we define "monotone entropy" and "sequential projections" and discuss some properties that simplify the computation of their actual numerical values.

III. PREVIOUS WORK

The aim of lossless quantum compression is to compress the unknown output $|\psi_j\rangle$ of an ensemble $\mathcal{E}=\{(p_i,|\psi_i\rangle)\}$ of quantum states using a variable-length quantum code so that the original state $|\psi_j\rangle$ can always be retrieved exactly and without error. When the $|\psi_i\rangle$'s are orthogonal, this is equivalent to lossless classical compression. The challenge is therefore to encode $\mathcal{E}$ when the $|\psi_i\rangle$'s are nonorthogonal and the code words might have indeterminate lengths.

In this section, we first give a definition of quantum bit strings that consist of a superposition of classical strings of different lengths. Then we describe previous work on how to use such quantum strings for compression and the difficulties that arise in such models. Finally, we outline a physical interpretation of these indeterminate-length quantum strings.

A. Indeterminate-length quantum bit strings

The strategy of classical variable-length compression is to assign short code words $C(x)$ to frequent events x (e.g., to frequent symbols in a text in some natural language), while rare events are assigned the remaining long code words. Trying a similar approach in quantum information theory naturally produces code words that are *superpositions of classical strings of different lengths*.

For example, suppose we have two letters A and B and a classical code C of the form $C(A)=0$ and $C(B)=11$. If, as a first naive try, we extend this map unitarily to quantum states spanned by $|A\rangle$ and $|B\rangle$, we get, for example,

$$C\left(\frac{|A\rangle+|B\rangle}{\sqrt{2}}\right)=\frac{|0\rangle+|11\rangle}{\sqrt{2}},$$

which does not have a determinate length, since it is in a superposition of lengths 1 and 2. It is called an *indeterminate-length quantum bit string*. Such strings can formally be defined as follows.

Definition III 1 (quantum bit string). A quantum state $|\psi\rangle$ is a quantum bit string (or qubit string) if it is an element of the Fock space (or string space)

$$\mathcal{H}_{\{0,1\}^*}:=\bigoplus_{n=0}^{\infty}(\mathbb{C}^2)^{\otimes n},$$

that is, if it can be expressed as a superposition of classical bit strings of the form

$$|\psi\rangle=\sum_{s\in\{0,1\}^*}\alpha_s|s\rangle,$$

with $\alpha_s\in\mathbb{C}$ and $\sum_{s\in\{0,1\}^*}|\alpha_s|^2=1$.

For convenience, we will sometimes drop the normalization condition. Moreover, it sometimes makes sense to call normal *mixed states*—i.e., density operators—on $\mathcal{H}_{\{0,1\}^*}$ qubit strings, too. The reason is that the prefixes of pure qubit strings can be mixed, which will be explained in detail below in Sec. V.

Boström and Felbinger [4] defined two ways to quantify the lengths of indeterminate-length strings.

Definition III 2 (length of qubit strings [4]). The base length L of an indeterminate-length string $|\psi\rangle=\sum_{s\in\{0,1\}^*}\alpha_s|s\rangle$ is the length of the longest part of its superposition,

$$L(\psi)=L\left(\sum_{s\in\{0,1\}^*}\alpha_s|s\rangle\right):=\max_{\alpha_s\neq 0}\ell(s),$$

or ∞ if the maximum does not exist. This can also be written as $L(\psi)=\max\{\ell(s)\,|\,\langle s|\psi\rangle\neq 0\}$. The average length $\bar{\ell}$ of an indeterminate-length quantum bit string is the expectation value of the length

$$\bar{\ell}(\psi)=\bar{\ell}\left(\sum_{s\in\{0,1\}^*}\alpha_s|s\rangle\right)=\sum_{s\in\{0,1\}^*}|\alpha_s|^2\ell(s),$$

which may as well be infinite. It can be written as $\bar{\ell}(\psi)=\langle\psi|\Lambda|\psi\rangle$, where Λ is the length operator, defined by linear extension of

$$\Lambda|s\rangle = \ell(s)|s\rangle \quad (s \in \{0,1\}^*).$$

Formally, Λ is an unbounded self-adjoint operator, defined on a dense subspace of $\mathcal{H}_{\{0,1\}^*}$.

If the length of a quantum string is observed, then $\bar{\ell}$ gives the expected length that is observed and L gives the maximum length that can be observed. However, given an unknown indeterminate-length string $|\psi\rangle$, neither its average length nor its base length can be measured without disturbing it.

B. Can indeterminate-length quantum strings be used for coding?

Various papers [3–7] have described problems in using indeterminate-length strings for lossless quantum data compression. Braunstein *et al.* [5] pointed out three difficulties of data compression with indeterminate-length strings. The first is that if the indeterminate-length strings are unknown to both the sender and the receiver, then it becomes impossible to synchronize the different computational paths (taking different numbers of time steps) that are performed on the strings.

The second difficulty is that if a mixture of indeterminate-length strings is transmitted at a fixed speed, then the recipient can never be sure when a message has arrived and the strings can be decompressed. The third difficulty is that if the data compression is performed by a read-write head (like a Turing machine), then after the data compression, the head location of the sender is entangled with the "lengths" of the indeterminate-length string which represents the compressed data.

Koashi and Imoto [6] argued that it is impossible to faithfully encode a mixture of nonorthogonal quantum states if the particular output states of the quantum information source *are not known*. They modeled lossless data compression as taking place in a register of N qubits. A compressed state in the register would be an unknown indeterminate-length quantum string with base length L, in which case only the remaining $N-L$ qubits would be usable by other applications without disturbing the compressed state. However, the base length L is not an observable; thus, the other applications cannot determine how many qubits are available. Thus the remaining $N-L$ qubits are not available for other applications to use, unless there is some *a priori* knowledge about L for some reason.

Schumacher and Westmoreland [7] showed that lossless quantum compression cannot be carried out by a unitary operation in a simple model of communication. They envisaged that indeterminate-length quantum strings would be padded with zeros to create determinate length strings (we explain this in more detail below in Sec. III C). They modeled the data compression as taking place between two parties Alice and Bob in which Alice sends Bob only the original strings (with the zero-padding removed), leaving Alice with a number of zeros depending on the length of the string she sent. If she sends Bob an indeterminate-length string, then after the transmission, Alice and Bob are entangled by the number of zeros that are left on Alice's register.

Boström and Felbinger [4] argued that it is not useful to consider quantum generalizations of classical prefix-free codes: classical prefix-free strings carry their own length information, but the length information in an indeterminate-length quantum string cannot be observed without disturbing the string. Their solution to this problem is to use a classical side channel to inform the receiver where to separate the code words.

Ahlswede and Cai [3] followed the same idea by sending the length information over a classical side channel. Compared to Ref. [4], they improved the compression rate by giving a more efficient way to use the side channel and they characterized the optimal compression rate in this setting. We describe both approaches in more detail below in Sec. III D.

However, in both cases, the use of the classical side channel requires that the sender (Alice) know the output of the quantum information source (at least partially) and thus the length of the compressed code word. This is in contrast to the situation examined in this paper.

C. Schumacher and Westmoreland's prefix-free average length compression

Schumacher and Westmoreland [7] investigated the general properties of indeterminate-length strings. An indeterminate-length string can be padded with zeros such that it consists of a determinate number of qubits.

Definition III 3 (zero extended form). If $|\psi\rangle = \sum_{\ell(s)\leq l_{\max}} \alpha_s |s\rangle$ is a quantum string in a register of l_{max} qubits, then its zero-extended form is

$$|\psi_{\mathrm{zef}}\rangle = \sum_{\ell(s)\leq l_{max}} \alpha_s |s0^{\otimes l_{max}-\ell(s)}\rangle.$$

This string has a determinate length of l_{max} qubits.

Given a sequence of N strings, it is useful to be able to "concatenate" them so that the strings are packed together at the beginning of the string and the zero padding all lies at the end of the sequence. Schumacher and Westmoreland [7] call this the "condensation operation."

Definition III 4 (condensable strings [7]). A code is condensable if for every N, there exists a unitary operation U such that

$$U(|\psi^1_{\mathrm{zef}}\rangle \otimes \cdots \otimes |\psi^N_{\mathrm{zef}}\rangle) = (|\psi^1\rangle \otimes \cdots \otimes |\psi^N\rangle)_{\mathrm{zef}}$$

for all the code words $|\psi^i\rangle$ which are "length eigenvectors,"—i.e. if each $|\psi^i\rangle$ contains in its superposition only classical words of some fixed length.

For example, if $|\psi^1\rangle = |0\rangle$, $|\psi^2\rangle = |10\rangle$, and $|\psi^3\rangle = |111\rangle$, then the condensation operation U is

$$\begin{aligned} U(|\psi^1_{\mathrm{zef}}\rangle \otimes |\psi^2_{\mathrm{zef}}\rangle \otimes |\psi^3_{\mathrm{zef}}\rangle) &= U(|000\rangle \otimes |100\rangle \otimes |111\rangle) \\ &= |0\rangle \otimes |10\rangle \otimes |111\rangle \otimes |000\rangle \\ &= (|0\rangle \otimes |10\rangle \otimes |111\rangle)_{\mathrm{zef}}. \end{aligned}$$

Superpositions of classical prefix-free strings are condensable. More generally, Schumacher and Westmoreland gave a definition of prefix-free quantum strings and showed that they are condensable. According to their definition, two strings are prefix free if when the additional qubits in the longer string are traced out, the resulting prefixes are orthogonal.

Definition III 5 (zero-padded prefix freedom [7]). Suppose $|\psi\rangle$ and $|\varphi\rangle$ are quantum strings with $n := L(|\psi\rangle) < L(|\varphi\rangle)$ and that they are in a register of l_{max} qubits. The first n qubits of $|\varphi_{\text{zef}}\rangle$ may be in a mixed state, described by the density operator

$$\rho^{1\ldots n} = \text{Tr}_{n+1,\ldots,l_{\text{max}}}(|\varphi_{\text{zef}}\rangle\langle\varphi_{\text{zef}}|).$$

The strings $|\psi\rangle$ and $|\varphi\rangle$ are prefix free if

$$\langle\psi|\rho^{1,\ldots,n}|\psi\rangle = 0.$$

This definition assumes that the two strings have determinate length. Two of the authors defined prefix-free strings more generally such that they can be supported on subspaces which are spanned by indeterminate-length quantum bit strings [8]. We give a review of this more general definition in Sec. V, which in fact contains Definition III 5 as a theorem (Lemma V 5).

Given many copies $\mathcal{E}^{\otimes n}$ of a quantum information source $\mathcal{E}$, Schumacher and Westmoreland further showed how to use prefix-free quantum bit strings for lossless compression (this corresponds to "step 1" of the compression process as described in the Introduction) using appropriate unitary operations. The indeterminate-length output is then projected onto the first $n[S(\mathcal{E})+\delta]$ qubits ("step 2"), where S is the von Neumann entropy. This projection (or partial trace) introduces only a small error which vanishes in the asymptotic case $n \to \infty$.

This can be seen as follows: Let ρ be the density operator corresponding to $\mathcal{E}$, with spectral decomposition

$$\rho = \sum_i p_i |i\rangle\langle i|.$$

Then $\mathcal{E}$ can be compressed by encoding each $|i\rangle$ as a prefix-free string of length $\lceil -\log_2(p_i) \rceil$ with zero padding. $\rho^{\otimes n}$ can be compressed in the same fashion by encoding every factor individually and applying the condensation operation to the resulting code words. Every string $|\psi\rangle$ in the typical subspace of $\rho^{\otimes n}$ has probability $\langle\psi|\rho|\psi\rangle$ arbitrarily close to $2^{-nS(\rho)}$ as n grows large. Thus, vector states in the typical subspace of ρ are encoded in a classical manner as strings of length arbitrarily close to $nS(\rho)$. The image of the projection on the first $n[S(\rho)+\delta]$ qubits thus contains this typical subspace. As with overwhelming probability, the output is very close to this subspace, it can afterwards be decoded with high (but not perfect) fidelity.

Hence this compression scheme consists of two parts as already mentioned in the Introduction: in a first step, the quantum message is compressed losslessly, in the sense that the output has minimal *average length* of about $nS(\mathcal{E})$. In a second part, some "cutoff" takes place, introducing some small error, but preparing the output to be transmitted over conventional channels by transforming it to fixed length [9].

One of the results of this paper is to show how the first step can be accomplished to minimize the expected *base length* (in the case of a single source), thus complementing the work by Schumacher and Westmoreland.

D. Compression with classical side channels

Boström and Felbinger [4] gave a scheme for lossless quantum compression of *known* ensemble outputs using classical side channels. If $\mathcal{E}=\{p_i,|\psi_i\rangle\}$ is the mixture to be compressed, then they assume that the value of i is known to the compressor, Alice. If she encodes $\mathcal{E}$ using a unitary operation C, then she sends the base length of the compressed string to Bob, the decompressor, through a classical side channel. She then sends $L(C(|\psi_i\rangle))$ qubits of $|\psi_i\rangle$'s zero-extended form to Bob. Since the length of the encoded string is encoded classically, it is not necessary to use a prefix-free code to encode the quantum part—thus C is unitary, but not necessarily a condensation operation.

Ahlswede and Cai [3] studied quantum data compression with classical side channels in more detail. They found an expression for the number of qubits that are sent through the quantum channel in Boström and Felbinger's lossless quantum compression scheme [4].

Moreover, they showed by using counterexamples that the optimal rate of compression R of a one-one code cannot be achieved by a greedy algorithm. However, the main goal of Ahlswede and Cai was to find a more efficient way to use the classical side channel than just to report the base lengths. They showed that the quantum part could be compressed further than in the scheme set out by Boström and Felbinger.

The basis for their scheme is as follows. If $\mathcal{E}$ is the mixture to be compressed and if there exists some small subspace X such that several states $|\psi_i\rangle$ lie exactly within X, then this fact can be reported through the classical side channel. Thus the amount of quantum information that must be sent through the quantum channel is reduced. They gave an expression for the optimal rate of compression in their scheme of an ensemble $\mathcal{E}=\{p_i,|\psi_i\rangle\}$ when the states $|\psi_i\rangle$ are linearly independent (but not necessarily orthogonal).

Compression with classical side channels has been studied in more detail for lossy compression [10]. Hayashi and Matsumoto [11] investigated variable- (but not indeterminate-) length universal compression.

Rallan and Vedral [12] gave another scheme for lossless quantum compression with classical side channels which does not use zero-extended forms. They envisaged that the compressed state would be represented by photons—thus using a tertiary alphabet $\{|0\rangle,|1\rangle,|\#\rangle\}$, where $|\#\rangle$ denotes the absence of a photon and marks the end of the string. They assumed that Alice has n copies of an ensemble $\mathcal{E}$ which she would like to send to Bob. In this scheme, Alice only sends Bob the value of n through the classical channel. This scheme has a nice physical interpretation.

E. Physical interpretation of indeterminate-length strings

Boström and Felbinger [4] pointed out that variable-length quantum strings can be realised in a quantum system whose particle number is not conserved. Rallan and Vedral [12] described in detail an example system where the average length of a string can be interpreted as its energy.

A Hilbert space $H^{\otimes n}$ can be realized by a sequence of photons $|\phi_1\rangle \otimes \cdots \otimes |\phi_n\rangle$ in which $|\phi_i\rangle$ represents exactly one photon with frequency ω_i. The value of the qubit $|\phi_i\rangle$ is

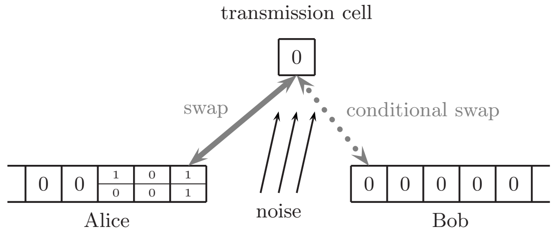

FIG. 1. Schematic of an always-open channel as described in Sec. IV.

realized by the polarization of its photon, either horizontal $|0\rangle$ or vertical $|1\rangle$. The absence of a photon at a particular frequency can be represented by $|\#\rangle$, which is orthogonal to $|0\rangle$ and $|1\rangle$. Indeterminate-length strings are obtained by allowing the number of photons to exist in superposition and ordering the photons by their frequencies. The first $|\#\rangle$ (which can be in a superposition of positions) is used to mark the end of the string.

The frequency of each photon $|\phi_i\rangle$ is chosen to be approximately equal so that $\omega_i \approx \omega$ for some value ω. The energy in a superposition of photons is the average energy required to either create or destroy that superposition ($\hbar\omega$ per photon of frequency ω where $\hbar$ is Planck's constant). Thus the energy of an indeterminate-length string of photons $|\phi\rangle$ is proportional to its average length and is given by (approximately) $\hbar\omega\bar{\ell}(|\phi\rangle)$.

On the other hand, the base length of $|\phi\rangle$ represents the number of photons at different frequencies that are used to describe $|\phi\rangle$. Thus the base length of $|\phi\rangle$ is the size of the system required to carry the state $|\phi\rangle$.

IV. COMMUNICATION MODEL FOR LOSSLESS QUANTUM COMPRESSION

Now we describe a model of a communication channel where lossless quantum compression of unknown mixtures is possible and useful.

The main argument why lossless quantum compression of unknown states seems to be impossible is that it is impossible to determine how many qubits to transmit when the message is in a superposition of different lengths. If Alice has an unknown indeterminate-length qubit string, how can she find out when the transmission is finished and the channel can be closed? To avoid this problem, we look at *always-open channels*.

A model of an always-open channel [13,14] is shown in Fig. 1. Suppose Alice wants to send Bob a single code word of a quantum prefix code—i.e., an indeterminate-length qubit string $|\psi\rangle$ which is an element of a prefix-free subspace $\mathcal{H}$ of $\mathcal{H}_{\{0,1\}^*}$ that Alice and Bob have agreed upon in advance. (This single code word might itself be a concatenation of several prefix-free code words.) As we shall see later in Theorem VI 4, we may assume that $\mathcal{H}$ is spanned by the classical code words of a classical prefix code as in Sec. III C above.

The main part of the channel is a transmission cell which carries exactly one qubit. Initially, this qubit is set to zero and so are all the qubits in Bob's memory. Moreover, Alice's memory contains a zero-padded form of her message string, as described in Definition III 3.

Now we describe the communication protocol—for each step, we describe what Alice and Bob are doing in the case of classical bits (i.e., in the case that the message qubit string $|\psi\rangle$ is just a classical string $|s\rangle$ out of the classical prefix-free orthonormal basis of $\mathcal{H}$), and we assume that the resulting operation is linearly extended to a unitary operation on the corresponding quantum system. The unitarity of the operations at Alice's and Bob's side is then assured by the reversibility of the corresponding classical operations.

At step i of the transmission, Alice swaps the ith qubit of her padded message string with the content of the transmission cell. Afterwards, Bob checks if the $i-1$ qubits he has received previously form a valid code word or not. (Due to prefix freedom, if the answer is "yes," then the transmission must be over—cf. also Definition III 5 and Lemma V 5). If the answer is "no," he swaps the ith qubit of his memory (which is just a zero) with the content of the transmission cell; otherwise, he does not do anything. That is, Bob applies a conditional swap, where the condition is that the transmission is not yet finished.

This way, the message qubit string is transmitted qubit by qubit. In the end, Alice ends up with a memory full of zeros, while the transmission cell contains a zero as well, and Bob's memory carries the zero-padded message string. Hence the entanglement problem described by Schumacher and Westmoreland [7] is avoided, and the message qubit string is transmitted reversibly and unitarily from Alice to Bob.

But what is the advantage of compression for such a communication channel if that channel can never be switched off by Alice or Bob? It cannot be used to save transmission time (considered as a resource), because both parties never know if the transmission is already finished or not (unless some predefined maximal transmission time t_{max} has passed). However, quantum compression can have other advantages: for example, suppose the transmission cell is subject to noise during the transmission. That is, the transmission of every single qubit has an inherent error probability. In this case, Alice can minimize transmission errors by compressing her quantum messages before sending them.

To be more exact, as soon as the code word has been fully transmitted (i.e., at a time step corresponding to the message's base length), Bob stops to access the transmission cell. Thus, any noise that affects the cell from that point on will not disturb the communication anymore, because the channel to Bob is effectively closed. Thus, minimizing the number of qubits to be transmitted reduces the probability of transmission errors, even though neither Alice nor Bob knows the number of transmitted qubits.

It is clear that the optimal compression method depends on the kind of noise that the system is exposed to. Obviously, in the case that each transmitted qubit is independently subject to the same kind of perturbation, then Schumacher and Westmoreland's average length compression method optimally minimizes transmission errors. But there are other conceivable scenarios: for example, we might have several channels at once that are subject to the same kind of noise or time-dependent noise that grows with the number of qubits.

In this case, it is not so clear any more what the best method of compression is.

In this paper, we compute the best possible compression method and the rate for minimizing the code's expected base length. Although we do not currently know of a natural noise model where the expected base length determines the error probability, it seems likely that there are indeed natural situations where this kind of compression is superior to average length compression—for example, models like those mentioned in the last paragraph where "later" qubits are subject to larger errors than "earlier" ones.

V. PREFIX-FREE QUANTUM BIT STRINGS

Schumacher and Westmoreland [7] defined prefix-free quantum strings in terms of their zero-extended forms using the partial trace; see Definition III 5 above. In Ref. [8], two of the authors have given another way to define prefix-free quantum strings which is more general and a more direct generalization of the classical definition. It can be shown to contain the definition by Schumacher and Westmoreland as a special case. In this section, we briefly review the definition and basic results on prefix-free quantum strings.

The notion of the prefix of a classical string is closely related to the concatenation operation $\circ$. Thus, before we define prefix-free quantum strings, we first explain how to concatenate quantum bit strings. If $|\psi\rangle\in\mathcal{H}_{\{0,1\}^*}$ is any quantum bit string and $s\in\{0,1\}^*$ is a classical bit strings, then we can define $|\psi\rangle\circ s$ by linear extension of the classical concatenation: Expand $|\psi\rangle=\sum_{x\in\{0,1\}^*}\alpha_x|x\rangle$, and define

$$|\psi\circ s\rangle := |\psi\rangle\circ s := \sum_{x\in s}\alpha_x|x\circ s\rangle.$$

Moreover, if $|\varphi\rangle=\sum_{t\in\{0,1\}^*}\beta_t|t\rangle$ is another qubit string with finite base length, then we set

$$|\psi\circ\varphi\rangle := |\psi\rangle\circ|\varphi\rangle := \sum_{t\in\{0,1\}^*}\beta_t|\psi\circ t\rangle.$$

This concatenation operation on the quantum strings is related to the tensor product: If $|\psi\rangle$ is a length eigenstate (i.e., an eigenvector of the length operator Λ), then $|\psi\rangle\circ|\varphi\rangle=|\psi\rangle\otimes|\varphi\rangle$. However, if $|\psi\rangle$ is not a length eigenstate, then the concatenation operation is not always an isometry and thus not always physically meaningful [8].

We can now define prefix-free sets of quantum strings (e.g., prefix-free subspaces of the string space $\mathcal{H}_{\{0,1\}^*}$) by direct generalization of the classical definition. Although there are several *a priori* possible generalizations, they all turn out to be equivalent (for a proof see Ref. [8]). To state them, we use the symbol λ for the empty string of length zero.

Definition V 1 (prefix-free sets of qubit strings). A set $M\subset\mathcal{H}_{\{0,1\}^*}$ of qubit strings is called *prefix free*, if one of the four following equivalent conditions holds:

(i) For every $|\varphi\rangle,|\psi\rangle\in M$ and classical string $s\in\{0,1\}^*\setminus\{\lambda\}$, it holds that $\langle\varphi|\psi\circ s\rangle=0$.

(ii) For every $|\varphi\rangle,|\psi\rangle\in M$ and qubit string $|\chi\rangle\perp|\lambda\rangle$, it holds that $\langle\varphi|\psi\circ\chi\rangle=0$.

(iii) For every $|\varphi\rangle,|\psi\rangle\in M$ and classical strings $s,t\in\{0,1\}^*$ with $s\neq t$, it holds that $\langle\varphi\circ t|\psi\circ s\rangle=0$.

(iv) For every $|\varphi\rangle,|\psi\rangle\in M$ and qubit strings $|\chi\rangle,|\tau\rangle\in\mathcal{H}_{\{0,1\}^*}$ with $|\chi\rangle\perp|\tau\rangle$, it holds that $\langle\varphi\circ\tau|\psi\circ\chi\rangle=0$.

The relevant case for quantum compression is that M is itself a closed subspace of string space $\mathcal{H}_{\{0,1\}^*}$. To prove prefix freedom of such a subspace, it is sufficient to prove this property for an arbitrary orthonormal basis [8].

Lemma V 2. A subspace $\mathcal{H}\subset\mathcal{H}_{\{0,1\}^*}$ is prefix free if and only if it has a prefix-free orthonormal basis. In this case, *every* orthonormal basis of $\mathcal{H}$ is prefix free.

Example V 3. The following subspace $\mathcal{H}\subset\mathcal{H}_{\{0,1\}^*}$ is prefix free:

$$\mathcal{H} := \mathrm{span}\left\{\frac{1}{\sqrt{2}}(|1\rangle+|01\rangle),\frac{1}{\sqrt{2}}(|10\rangle-|010\rangle)\right\}.$$

It is easily checked that condition (i) from Definition V 1 above is satisfied for the two orthonormal basis vectors.

Similarly as in the classical case, closed prefix-free subspaces obey a Kraft inequality [8].

Lemma V 4 (quantum Kraft inequality). Let $\{|e_i\rangle\}_{i\in I}\subset\mathcal{H}_{\{0,1\}^*}$ be a prefix-free orthonormal system, spanning a closed subspace $\mathcal{H}\subset\mathcal{H}_{\{0,1\}^*}$. Then, it holds that

$$\sum_{i\in I}2^{-L(e_i)}\leqslant\sum_{i\in I}2^{-\bar{\ell}(e_i)}\leqslant\mathrm{Tr}[2^{-\Lambda}\mathbb{P}(\mathcal{H})]\leqslant 1,$$

where $\mathbb{P}(\mathcal{H})$ denotes the orthogonal projector onto $\mathcal{H}$. Equality holds for the left three terms if and only if every $|e_i\rangle$ is a length eigenvector.

Prefix-free subspaces have a remarkable property: every basis vector of length n can be distinguished with certainty from every other (even longer) basis vector by measuring the first n qubits only. Unfortunately, this is only true in general for orthonormal bases of length eigenvectors [8].

Lemma V 5. An orthonormal system $M\subset\mathcal{H}_{\{0,1\}^*}$ which consists entirely of length eigenvectors is prefix free if and only if for every $|\varphi\rangle,|\psi\rangle\in M$ with $|\varphi\rangle\neq|\psi\rangle$, it holds that

$$\langle\psi|\varphi^{\ell(\psi)}|\psi\rangle=0, \tag{1}$$

where φ^n denotes the restriction of the quantum state $|\varphi\rangle\langle\varphi|$ to the first $n:=\ell(\psi)$ qubits.

This lemma shows that if the subspace contains an orthonormal basis of length eigenvectors, our definition of prefix freedom is equivalent to the definition by Schumacher and Westmoreland [7].

We have only collected the basic facts about prefix-free quantum bit strings that are relevant for lossless quantum compression. For more details, we refer the reader to Refs. [7,8].

In general, the concatenation operation does not preserve the norm of vectors from $\mathcal{H}_{\{0,1\}^*}$; i.e., it is not an isometry and hence not physically meaningful. However, we shall now prove that concatenation can be implemented in principle on a quantum computer (i.e., it is an isometry) if one restricts to prefix-free Hilbert spaces:

Theorem V 6 (isometry of concatenation). If $\{|\varphi_1\rangle,|\varphi_2\rangle\}\subset\mathcal{H}_{\{0,1\}^*}$ is a prefix-free set and $|\psi_1\rangle,|\psi_2\rangle\in\mathcal{H}_{\{0,1\}^*}$, then

$$\langle \varphi_1 \circ \psi_1 | \varphi_2 \circ \psi_2 \rangle = \langle \varphi_1 | \varphi_2 \rangle \langle \psi_1 | \psi_2 \rangle .$$

Consequently, if $\mathcal{H} \subset \mathcal{H}_{\{0,1\}^*}$ is a closed prefix-free subspace, then there exists a unique isometry $U_\circ : \mathcal{H} \otimes \mathcal{H}_{\{0,1\}^*} \to \mathcal{H}_{\{0,1\}^*}$ such that $U_\circ |\varphi\rangle \otimes |\psi\rangle = |\varphi \circ \psi\rangle$ for every $|\varphi\rangle \in \mathcal{H}$ and $|\psi\rangle \in \mathcal{H}_{\{0,1\}^*}$.

Note that in the special case that $\mathcal{H}$ is spanned by length eigenvectors, the map $U_\circ$ corresponds to the "simple condensation operation" as defined by Schumacher and Westmoreland [7].

Proof. It is easy to check that for every pair of qubit strings $|\varphi_1\rangle, |\varphi_2\rangle \in \mathcal{H}_{\{0,1\}^*}$ and $s \in \{0,1\}^*$, we have $\langle \varphi_1 \circ s | \varphi_2 \circ s \rangle = \langle \varphi_1 | \varphi_2 \rangle$. Now suppose that additionally $\Phi := \{|\varphi_1\rangle, |\varphi_2\rangle\}$ is a prefix-free set and $|\psi\rangle \in \mathcal{H}_{\{0,1\}^*}$ is an arbitrary qubit string. Expanding $|\psi\rangle = \sum_{s \in \{0,1\}^*} \gamma_s |s\rangle$, we have

$$\begin{aligned}\langle \varphi_1 \circ \psi | \varphi_2 \circ \psi \rangle &= \sum_{s,t \in \{0,1\}^*} \bar{\gamma}_s \gamma_t \langle \varphi_1 \circ s | \varphi_2 \circ t \rangle \\ &\overset{(*)}{=} \sum_{s \in \{0,1\}^*} \bar{\gamma}_s \gamma_s \langle \varphi_1 \circ s | \varphi_2 \circ s \rangle \\ &= \langle \varphi_1 | \varphi_2 \rangle \sum_{s \in \{0,1\}^*} |\gamma_s|^2 \\ &= \langle \varphi_1 | \varphi_2 \rangle \langle \psi | \psi \rangle .\end{aligned}$$

In the relation labeled with (*), we have used the fact that Φ is prefix free, and so $\langle \varphi_1 \circ s | \varphi_2 \circ t \rangle = 0$ if $s \neq t$. Finally, if $|\psi_1\rangle, |\psi_2\rangle \in \mathcal{H}_{\{0,1\}^*}$ are arbitrary qubit strings, then choose an arbitrary orthonormal basis $\{|e_i\rangle\}_{i \in \mathbb{N}}$ of $\mathcal{H}_{\{0,1\}^*}$ such that $|\psi_1\rangle = \lambda |e_1\rangle$ with $\lambda \in \mathbb{R}$ and expand $|\psi_2\rangle$ as $|\psi_2\rangle = \sum_{i \in \mathbb{N}} \alpha_i |e_i\rangle$. It follows that

$$\begin{aligned}\langle \varphi_1 \circ \psi_1 | \varphi_2 \circ \psi_2 \rangle &= \sum_{i \in \mathbb{N}} \alpha_i \langle \varphi_1 \circ \psi_1 | \varphi_2 \circ e_i \rangle \\ &\overset{(**)}{=} \alpha_1 \langle \varphi_1 \circ \psi_1 | \varphi_2 \circ e_1 \rangle \\ &= \alpha_1 \lambda \langle \varphi_1 | \varphi_2 \rangle \underbrace{\langle e_1 | e_1 \rangle}_{=1} \\ &= \langle \psi_1 | \psi_2 \rangle \langle \varphi_1 | \varphi_2 \rangle .\end{aligned}$$

In the relation labeled with (**), we have again used the fact that Φ is prefix free, and consequently $\langle \varphi_1 \circ \psi_1 | \varphi_2 \circ e_i \rangle = 0$ for $i \geq 2$, since $|\psi_1\rangle \perp |e_i\rangle$. ■

We show now that the base length of a concatenation of two-qubit strings is the sum of the individual base lengths. Note that this is in general not true for average length $\bar{\ell}$: for example, if $|\psi\rangle = \frac{1}{\sqrt{2}}(|1\rangle + |01\rangle)$ and $|\varphi\rangle = \frac{1}{\sqrt{2}}(|10\rangle - |010\rangle)$ are two vectors from the prefix-free Hilbert space $\mathcal{H}$ in Example V 3 and if $|\chi\rangle := \frac{1}{\sqrt{2}}(|\psi\rangle + |\varphi\rangle)$, then it is easy to check that $\frac{19}{4} = \bar{\ell}(\chi \circ \varphi) > \bar{\ell}(\chi) + \bar{\ell}(\varphi) = 2 + \frac{5}{2}$.

Lemma V 7 (additivity of base length). If $|\varphi\rangle, |\psi\rangle \in \mathcal{H}_{\{0,1\}^*}$ are qubit strings with finite base lengths—i.e., $L(\varphi) < \infty$ and $L(\psi) < \infty$—then $L(\varphi \circ \psi) = L(\varphi) + L(\psi)$.

Proof. For every $|\varphi\rangle \in \mathcal{H}_{\{0,1\}^*}$, define $S(\varphi) := \{ s \in \{0,1\}^* | \langle s | \varphi \rangle \neq 0 \}$. It follows that $L(\varphi) = \max\{\ell(s) | s \in S(\varphi)\}$. If we expand $|\varphi\rangle =: \sum_{s \in \{0,1\}^*} \alpha_s |s\rangle$ and $|\psi\rangle =: \sum_{t \in \{0,1\}^*} \beta_t |t\rangle$, then

$$|\varphi \circ \psi\rangle = \sum_{s,t} \alpha_s \beta_t |s \circ t\rangle .$$

It follows that

$$S(\varphi \circ \psi) \subseteq S(\varphi) \circ S(\psi) := \{ s \circ t | s \in S(\varphi), t \in S(\psi) \},$$

and thus

$$\begin{aligned} L(\varphi \circ \psi) &= \max\{\ell(s) | s \in S(\varphi \circ \psi)\} \\ &\leq \max\{\ell(s) | s \in S(\varphi) \circ S(\psi)\} \\ &= \max\{\ell(s \circ t) | s \in S(\varphi), t \in S(\psi)\} \\ &= \max_{s \in S(\varphi)} \ell(s) + \max_{t \in S(\psi)} \ell(t) = L(\varphi) + L(\psi). \end{aligned}$$

Let now s_{max} and t_{max} be elements of maximal length in $S(\varphi)$ and $S(\psi)$, respectively. Clearly, $\langle s_{max} \circ t_{max} | \varphi \circ \psi \rangle = \Sigma \alpha_s \beta_t$, where the sum is over all $s \in S(\varphi)$ and $t \in S(\psi)$ such that $s \circ t = s_{max} \circ t_{max}$. But because of the maximum-length property of s_{max} and t_{max}, it follows that $\ell(s) = \ell(s_{max})$ and $\ell(t) = \ell(t_{max})$, and thus $s = s_{max}$ and $t = t_{max}$. Consequently, $\langle s_{max} \circ t_{max} | \varphi \circ \psi \rangle = \alpha_{s_{max}} \beta_{t_{max}} \neq 0$, and $L(\varphi \circ \psi) \geq L(\varphi) + L(\psi)$. ■

We explain the meaning of these results for lossless quantum data compression below after Definition VI 1, the definition of a lossless quantum code.

VI. LOSSLESS QUANTUM DATA COMPRESSION

Our aim is to compute the best possible rate for compressing the unknown output of a single quantum information source, where the source is given by an ensemble $\mathcal{E} = \{p_i, |\psi_i\rangle\}_i$ of in general nonorthogonal quantum states $|\psi_i\rangle$ with probabilities $p_i > 0$. As motivated in the Introduction, we want to minimize the expected base length of the code and we want to use a prefix-free code to allow concatenation of code words in the case of several sources.

Definition VI 1 (lossless quantum code). Let $\mathcal{E} = \{p_i, |\psi_i\rangle\}_i$ be an ensemble of quantum states in a Hilbert space, with $\mathcal{H} := \text{span}\{|\psi_1\rangle, \ldots, |\psi_n\rangle\}$. A lossless code C is an isometric linear map from $\mathcal{H}$ into a closed prefix-free subspace $\mathcal{H}' \subset \mathcal{H}_{\{0,1\}^*}$.

The expected base length of compression of C is

$$E(L(C(\mathcal{E}))) = \sum_i p_i L(C(|\psi_i\rangle)). \tag{2}$$

C is optimal if for any other code C',

$$E(L(C(\mathcal{E}))) \leq E(L(C'(\mathcal{E}))).$$

Expression (2) defines the compression rate of the code as the expected base length of the encoding of the output of a *single* instance of the ensemble. What if we have *n copies* $\mathcal{E}^{\otimes n}$ of an ensemble $\mathcal{E}$; i.e., several output states are produced independently and identically distributed according to $\mathcal{E}$?

Suppose we have two different ensembles $\mathcal{E} = \{p_i, |\psi_i\rangle\}$ and $\mathcal{F} = \{q_j, |\varphi_j\rangle\}$, which have optimal codes $C_\mathcal{E}$ and $C_\mathcal{F}$, respectively. As the codes are prefix free, we may concatenate them to obtain a code $C_\mathcal{E} \circ C_\mathcal{F}$ for $\mathcal{E} \otimes \mathcal{F}$. Theorem V 6 proves that this concatenation can be done unitarily—i.e., can be implemented in principle on a quantum computer—and Lemma

V 7 tells us that the base lengths then just add up. Explicitly,

$$\begin{aligned}E(L(C_{\mathcal{E}}\circ C_{\mathcal{F}})) &= \sum_{ij} p_i q_j L(C_{\mathcal{E}}(|\psi_i\rangle)\circ C_{\mathcal{F}}(|\varphi_j\rangle))\\ &= \sum_{ij} p_i q_j (L(C_{\mathcal{E}}(|\psi_i\rangle)) + L(C_{\mathcal{F}}(|\varphi_i\rangle)))\\ &= E(L(C_{\mathcal{E}})) + E(L(C_{\mathcal{F}})).\end{aligned}$$

Thus $C_{\mathcal{E}}\circ C_{\mathcal{F}}$ is a code for $\mathcal{E}\otimes\mathcal{F}$ with the simple property that its rate is just the sum of the rates of the two codes. However, it is not necessarily optimal anymore. In fact, denoting the optimal compression rate of an ensemble $\mathcal{E}$ by $R(\mathcal{E})$, Theorem VI 4 below will show that if, e.g.,

$$\mathcal{E} := \left\{\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right),(|0\rangle,|1\rangle,|2\rangle)\right\},$$

where $|0\rangle$, $|1\rangle$, and $|2\rangle$ are three arbitrary orthonormal vectors, then $R(\mathcal{E})=\frac{5}{3}$, while $R(\mathcal{E}\otimes\mathcal{E})=\frac{29}{9}<2R(\mathcal{E})=\frac{30}{9}$. Hence concatenation of codes does not always produce optimal codes (although they are typically quite good), and our result will, for example, not give a simple expression for the asymptotic rate $\lim_{n\to\infty}\frac{1}{n}R(\mathcal{E}^{\otimes n})$, only the upper bound $R(\mathcal{E})$.

Yet the result is nevertheless useful, in particular if there is only one output of the source or if there are several sources $\mathcal{E}_1\otimes\mathcal{E}_2\otimes\cdots\otimes\mathcal{E}_k$, which are not known in advance to the compressor. Then, the compression can be done sequentially, for one source after the other, and the code words are concatenated while the rates just add up. As for the compression rate, we get the useful upper bound $R\leqslant\sum_{i=1}^{k}R(\mathcal{E}_i)$ even if there is no translation invariance in the sequence of sources.

This subadditivity property of the optimal rate also shows that in the case of n copies of one source, block coding with concatenation will produce the optimal asymptotic compression rate: write

$$\mathcal{E}^{\otimes n} = \mathcal{E}^{\otimes n_1}\otimes\mathcal{E}^{\otimes n_2}\otimes\cdots\otimes\mathcal{E}^{\otimes n_k},$$

with $\sum_{i=1}^{k}n_k=n$ such that the sequence $(n_k)_{k\in\mathbb{N}}$ is increasing. Then, use the optimal code C_{n_k} for each block $\mathcal{E}^{\otimes n_k}$ separately and concatenate the codes to get a code for $\mathcal{E}^{\otimes n}$. The corresponding compression rate will be asymptotically optimal.

To state the optimal compression rate for single sources, we introduce the notion of *monotone entropy* and of a *sequential projection* of some ensemble $\mathcal{E}=\{p_i,|\psi_i\rangle\}$.

Definition VI 2 (monotone entropy). Let $p=(p_1,p_2,\ldots,p_n)$ be a probability vector. Then, we define the monotone entropy $H_{mon}(p)$ as

$$H_{mon}(p) := \min\left\{\sum_{i=1}^{n}p_i\ell_i \,\middle|\, \sum_{i=1}^{n}2^{-\ell_i}\leqslant 1,\ \underbrace{\ell_1\leqslant\ell_2\leqslant\cdots\leqslant\ell_n}_{(*)},\ell_i\in\mathbb{N}_0\right\}.$$

Note that the Kraft inequality on the right-hand side implies that the values $\{\ell_i\}_i$ are code word lengths of a prefix code.

Suppose we removed $(*)$ from the definition. This would mean that we look for the smallest possible rate of any prefix code for the given probability distribution p. As is well known, this best rate is given by the Shannon entropy $H(p)$; thus, we would get back (up to possibly one bit) Shannon entropy. This implies

$$H_{mon}(p)\geqslant H(p) \tag{3}$$

and justifies that we call H_{mon} an *entropy*. Note that H_{mon} changes if we permute the entries of p (while Shannon entropy stays constant). If the elements of p are in decreasing order, then monotone entropy equals Shannon entropy up to possibly one bit:

$$p_1\geqslant p_2\geqslant\cdots\geqslant p_n\Rightarrow H(p)\leqslant H_{mon}(p)\leqslant H(p)+1. \tag{4}$$

This is easily proved by inserting $\ell_i:=\lceil -\log_2 p_i\rceil$. On the other hand, if we set $\ell_i:=\lceil\log_2 n\rceil$ for every i, we get the universal upper bound

$$H_{mon}(p)\leqslant\lceil\log_2 n\rceil, \tag{5}$$

if n denotes the number of elements in p.

Now we explain the notion of a *sequential projection*. It is a certain probability distribution which is constructed from $\mathcal{E}$ in a sequential manner.

Definition VI 3 (sequential projection). Let $\mathcal{E}=\{(p_i,|\psi_i\rangle)\}_{i=1}^{n}$ be an ensemble of quantum states. A sequential projection $p'=(p'_1,p'_2,\ldots,p'_k)$ is any probability distribution which can be constructed by the following algorithm.

(i) Choose an arbitary integer $i_1\in\{1,\ldots,n\}$. Then, add up all the probabilities p_j that correspond to vectors $|\psi_j\rangle$ which are linearly dependent on (parallel to) $|\psi_{i_1}\rangle$ to get the value p'_1—i.e.,

$$I_1 := \{j\in\{1,\ldots,n\}\,|\,|\psi_j\rangle\in\mathrm{span}\{|\psi_{i_1}\rangle\}\}$$

(in particular, $i_1\in J$) and $p'_1:=\sum_{j\in I_1}p_j$.

(ii) Choose an arbitrary remaining integer $i_2\in\{1,\ldots,n\}\backslash I_1$. Add up all the probabilities p_j that correspond to vectors $|\psi_j\rangle$ which are linearly dependent on $|\psi_{i_2}\rangle$ and the previously chosen vectors in I_1 to get the value p'_2—i.e.,

$$I_2 := \{j\in\{1,\ldots,n\}\setminus I_1\,|\,|\psi_j\rangle\in\mathrm{span}(\{|\psi_{i_2}\rangle\}\cup I_1)\}$$

and $p'_2:=\sum_{j\in I_2}p_j$.

(iii) Choose an arbitrary remaining integer $i_3\in\{1,\ldots,n\}\backslash(I_1\cup I_2)$. Add up all the probabilities p_j that correspond to vectors $|\psi_j\rangle$ which are linearly dependent on $|\psi_{i_3}\rangle$ and the previously chosen vectors to get the value p'_3—i.e.,

$$I_3 := \{j\in\{1,\ldots,n\}\setminus(I_1\cup I_2)\,|\,|\psi_j\rangle\in\mathrm{span}(\{|\psi_{i_3}\rangle\}\cup I_1\cup I_2)\}$$

and $p'_3:=\sum_{j\in I_3}p_j$.

(iv) …

(v) Iterate these steps until there are no remaining vectors in the ensemble.

As an example of a sequential projection, consider the states from an ensemble $\{p_i,|\psi_i\rangle\}_{i=1}^{4}$:

$$|\psi_1\rangle=|0\rangle,\quad |\psi_2\rangle=|+\rangle=\frac{|0\rangle+|1\rangle}{\sqrt{2}},$$

$$|\psi_3\rangle = |1\rangle, \quad |\psi_4\rangle = |2\rangle,$$

where $|0\rangle$, $|1\rangle$, and $|2\rangle$ denote orthonormal basis vectors from an arbitrary Hilbert space. Then, applying the definition above and noting that $|\psi_3\rangle$ is in the span of $|\psi_1\rangle$ and $|\psi_2\rangle$, we get one possible sequential projection as

$$p_1' = p_1, \quad p_2' = p_2 + p_3, \quad p_3' = p_4,$$

where we have chosen $i_1=1$, $i_2=2$, and $i_3=4$. Other choices of indices yield different sequential projections. That is, to every ensemble $\mathcal{E}$, there are several possible sequential projections of $\mathcal{E}$. By combinatorics, the number of sequential projections to an ensemble $\mathcal{E}$ of n elements is upper bounded by $n!$. Each sequential projection is a probability vector with $\dim \mathcal{E}$ elements.

To get an idea how sequential projections are related to base length compression, suppose a code Q compresses the state $|0\rangle$ with base length l_1 and the state $|+\rangle$ with base length $l_2 \geq l_1$; then, since $|1\rangle$ is on the span of $|0\rangle$ and $|+\rangle$, $|1\rangle$ will typically also be compressed to l_2. Suppose $|2\rangle$ is compressed to length l_3 (which can safely be achieved if $l_3 \geq l_2$); then, the compression rate of $\mathcal{E}$ is

$$E(L(Q(\mathcal{E}))) = p_1 l_1 + (p_2 + p_3) l_2 + p_4 l_3 = p_1' l_1 + p_2' l_2 + p_3' l_3.$$

We can now state the optimal rate of compression in terms of monotone entropy and sequential projection, which can both be calculated combinatorially.

Theorem VI 4 (optimal compression rate). Let $\mathcal{E} = \{p_i, |\psi_i\rangle\}$ be an ensemble of quantum states in some Hilbert space. Then the optimal base length lossless quantum prefix compression code C can be constructed such that it maps into a Hilbert space $\mathcal{H}'$ which is spanned by an orthonormal basis of length eigenvectors. The rate R of this optimal code is given by

$$R = \min\{H_{mon}(p') | p' \text{ is a sequential projection of } \mathcal{E}\}.$$

In particular, if the vectors $\{|\psi_i\rangle\}_i$ are linearly independent, then $H(p) \leq R \leq H(p)+1$; i.e., the rate is essentially given by the Shannon entropy of $\mathcal{E}$'s probability distribution. In any case, we have the upper bound $R \leq H(p)+1$.

Before we give a proof, we illustrate the theorem with one example. Suppose our ensemble consists of eight states $|\psi_1\rangle, |\psi_2\rangle, \ldots, |\psi_8\rangle$ from some Hilbert space, each with probability $p_1 = p_2 = \cdots = p_8 = \frac{1}{8}$, such that the span of those eight states has dimension 4. Furthermore, suppose that any four of those states are linearly independent.

Our theorem tells us that we can compress the ensemble at least as good as $R \leq H(p)+1 = H(\frac{1}{8}, \frac{1}{8}, \ldots, \frac{1}{8})+1 = 4$, but we can do better than that. To compute the optimal compression rate, we have to look at all possible sequential projections.

We construct a sequential projection p': first, we arbitrarily choose one of the vectors—say, $|\psi_1\rangle$. As there is no other vector which is linearly dependent on (parallel to) $|\psi_1\rangle$, the first entry to p' is $p_1' := p_1 = \frac{1}{8}$.

As the second step, we choose one of the remaining vectors—say, $|\psi_2\rangle$. We have to check if there are any remaining vectors that are in the span of $|\psi_1\rangle$ and $|\psi_2\rangle$ (i.e., linearly dependent on those two), which is by assumption not the case. Thus, we get $p_2' := p_2 = \frac{1}{8}$.

We go on by choosing the next vector arbitrarily—say, $|\psi_3\rangle$. As there are no remaining vectors in the span of $|\psi_1\rangle$, $|\psi_2\rangle$, and $|\psi_3\rangle$, we also get $p_3' := p_3 = \frac{1}{8}$.

Then we select another remaining vector—say, $|\psi_4\rangle$. But now, all the remaining vectors $|\psi_5\rangle$, $|\psi_6\rangle$, $|\psi_7\rangle$, and $|\psi_8\rangle$, are in the linear span of $|\psi_1\rangle$, $|\psi_2\rangle$, $|\psi_3\rangle$, and $|\psi_4\rangle$. Thus, we have to add the corresponding probabilities to get $p_4' := p_4 + p_5 + p_6 + p_7 + p_8 = \frac{5}{8}$. Thus,

$$p' = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{5}{8}\right).$$

In this example, repeating the process with different choices of vectors will always result in the same probability distribution p'. Thus, in this case, there is only one possible sequential projection of $\mathcal{E}$, which is given above. We get the rate R by computing $R = H_{mon}(p')$. First, we know from (5) that $H_{mon}(p') \leq \lceil \log_2 4 \rceil = 2$. In fact, with the help of a little computer program, it is easy to see that the minimum in the definition of H_{mon} is indeed attained at this value—that is,

$$R = H_{mon}(p') = 2.$$

Now we prove this theorem.

Proof. The proof consists of two parts: first, we show that a rate of R is achievable, then we show that this rate is optimal. We shall denote our ensemble by

$$\mathcal{E} = \{\underbrace{p(|\psi_i\rangle)}_{=:p_i}, |\psi_i\rangle\}_{i=1}^{n},$$

and we write $\mathcal{H} := \text{span}\{|\psi_1\rangle, \ldots, |\psi_n\rangle\}$. For sequential projections, we use the nomenclature from Definition VI 3.

To see the achievability, let $p' = (p_1', \ldots, p_d')$ be an arbitrary sequential projection of $\mathcal{E}$. Let $(c_1, \ldots, c_d) \subset \{0,1\}^*$ be a prefix code with code word lengths $\ell_i := \ell(c_i)$ which are minimizers in the definition of $H_{mon}(p')$ (such a code exists due to the Kraft inequality). Let $\mathcal{H}' := \text{span}\{|c_1\rangle, \ldots, |c_d\rangle\} \subset \mathcal{H}_{\{0,1\}^*}$. We will now construct a code (a linear isometric map) $C: \mathcal{H} \to \mathcal{H}'$. For $i \in \{1, \ldots, d\}$, let Ψ_i be the set of vectors from $\mathcal{E}$ that have been chosen in step i of the construction of p', such that $\Psi_i \cap \Psi_j = 0$ for $i \neq j$, $\cup_{i=1}^{d} \Psi_i = \{|\psi_1\rangle, \ldots, |\psi_n\rangle\}$, and $p_i' = \Sigma_{|\psi\rangle \in \Psi_i} p(|\psi\rangle)$.

We start by specifying the action of C on the vectors of Ψ_1. All the vectors in Ψ_1 are equal up to some phase factor; i.e., they are equal to $e^{i\theta}|\psi\rangle$, where $|\psi\rangle \in \Psi_1$. We set

$$C|\psi\rangle := |c_1\rangle; \tag{6}$$

then, $L(C(|\psi\rangle)) = \ell(c_1)$ for every $|\psi\rangle \in \Psi_1$. Since dim span $(\Psi_1 \cup \Psi_2) = 2$, we can construct C such that

$$C(\text{span}(\Psi_1 \cup \Psi_2)) = \text{span}\{|c_1\rangle, |c_2\rangle\} \tag{7}$$

isometrically, while still respecting (6). Consequently, $L(C(|\psi\rangle)) = \max\{\ell(c_1), \ell(c_2)\} = \ell(c_2)$ for every $|\psi\rangle \in \Psi_2$ [here we use the monotonicity property $\ell(c_1) \leq \ell(c_2) \leq \cdots$]. The next step is to demand

$$C(\text{span}(\Psi_1 \cup \Psi_2 \cup \Psi_3)) = \text{span}\{|c_1\rangle, |c_2\rangle, |c_3\rangle\}$$

isometrically, while respecting (6) and (7). Iterating this process, we obtain a code C in the sense of Definition VI 1. The expected base length compression rate is

$$
\begin{aligned}
r &:= \sum_{j=1}^{n} p_j L(C(|\psi_j\rangle)) \\
&= \sum_{i=1}^{d} \left(\sum_{|\psi\rangle \in \Psi_i} p(|\psi\rangle) \right) \ell(c_i) \\
&= \sum_{i=1}^{d} p'_i \ell_i = H_{mon}(p').
\end{aligned}
$$

Next, we show that this code is optimal; i.e., no lossless code can beat monotone entropy. Thus, let $C:\mathcal{H}\to\mathcal{H}'$ be a code in the sense of Definition VI 1, and let $r(C) := \sum_i p_i L(C(|\psi_i\rangle))$ be the corresponding compression rate. We may assume that the vectors $|\psi_i\rangle$ are ordered such that $i < j \Rightarrow L(C(|\psi_i\rangle)) \leq L(C(|\psi_j\rangle))$.

We will now construct a sequential projection p' which corresponds to this code C. Let $i_1 := 1$ and $l_1 := L(C(|\psi_{i_1}\rangle))$. Suppose $|\psi_j\rangle \in I_1$; then, $|\psi_j\rangle$ is linearly dependent on $|\psi_{i_1}\rangle$. Since C is isometric, $C(|\psi_j\rangle)$ must be linearly dependent on $C(|\psi_{i_1}\rangle)$ as well, and so $L(C(|\psi_j\rangle)) = l_1$. So

$$\sum_{j\in I_1} p_j L(C(|\psi_j\rangle)) = \left(\sum_{j \in I_1} p_j \right) l_1 = p'_1 l_1.$$

Then, let i_2 be the smallest natural number which is not in I_1 and let $l_2 := L(C(|\psi_{i_2}\rangle))$. If $|\psi_j\rangle \in I_2$, then $|\psi_j\rangle$ is in the linear span of I_1 and $|\psi_{i_2}\rangle$. Since C is isometric, we can again conclude that $L(C(|\psi_j\rangle)) \leq l_2$. But if we had $L(C(|\psi_j\rangle)) < l_2 = L(C(|\psi_{i_2}\rangle))$, then it would follows that $j < i_2$, which is impossible. Hence $L(C(|\psi_j\rangle)) = l_2$ and

$$\sum_{j\in I_2} p_j L(C(|\psi_j\rangle)) = \left(\sum_{j \in I_2} p_j \right) l_2 = p'_2 l_2.$$

We iterate this procedure until all the vectors from the ensemble have been used. Since the vectors $|\psi_i\rangle$ are ordered according to their lengths, we have $l_1 \leq l_2 \leq l_3 \leq \cdots$ and so on. Moreover, these code word lengths satisfy the Kraft inequality. To see this, note that the vectors $|\psi_{i_k}\rangle$ are linearly independent and span the Hilbert space $\mathcal{H}'$. Let $\{|\varphi_k\rangle\}_k$ be the orthonormal basis of $\mathcal{H}'$ which is generated by the Gram-Schmidt orthonormalization process from the basis $\{|\psi_{i_k}\rangle\}_k$. It follows that

$$L(|\varphi_k\rangle) \leq \max_{k' \leq k} L(C(|\psi_{i_{k'}}\rangle)) = L(C(|\psi_{i_k}\rangle)) = l_k.$$

Since $\mathcal{H}'$ is a prefix Hilbert space, the quantum Kraft inequality from Lemma V 4 yields

$$\sum_k 2^{-l_k} \leq \sum_k 2^{-L(|\varphi_k\rangle)} \leq 1. \tag{8}$$

Moreover, $p' = (p'_1, p'_2, \ldots)$ is by construction a sequential projection. Hence

$$r(C) = \sum_{j=1}^{n} p_j L(C(|\psi_j\rangle)) = \sum_k p'_k l_k \geq H_{mon}(p'),$$

which concludes the optimality part of the proof. An easy additional argument shows that the optimal code Hilbert space $\mathcal{H}'$ may always be chosen to be spanned by an orthonormal basis of length eigenstates: Due to (8), there is a classical prefix-free code $\{c_k\}_k$ with $\ell(c_k) = L(|\varphi_k\rangle)$. Let $\mathcal{H}'' := \mathrm{span}_k |c_k\rangle$; then, $\mathcal{H}''$ is prefix free. Let $U|\varphi_k\rangle := |c_k\rangle$; then, U maps $\mathcal{H}'$ unitarily onto $\mathcal{H}''$. Hence, the composition $U \circ C$ is a lossless quantum code. Suppose $j \in I_k$; then, $|\psi_j\rangle \in \mathrm{span}_{k' \leq k} |\psi_{i_{k'}}\rangle$, and hence

$$U \circ C(|\psi_j\rangle) \in \mathrm{span}_{k'\leq k}\, U \circ C(|\psi_{i_{k'}}\rangle) = \mathrm{span}_{k' \leq k}\, U|\varphi_{k'}\rangle,$$

and so

$$
\begin{aligned}
L(U \circ C(|\psi_j\rangle)) &\leq \max_{k' \leq k} L(U|\varphi_{k'}\rangle) = \max_{k' \leq k} \ell(c_k) = \max_{k' \leq k} L(|\varphi_{k'}\rangle) \\
&\leq \max_{k' \leq k} L(|\psi_{i_{k'}}\rangle) = \max_{k' \leq k} l_{k'} = l_k = L(C(|\psi_j\rangle)).
\end{aligned}
$$

Thus, $r(U \circ C) \leq r(C)$; i.e., $U \circ C$ compresses at least as good as C.

In the special case that all the vectors $|\psi_i\rangle$ are linearly independent, the sequential projections of $\mathcal{E}$ are exactly the permutations of the probability distribution p. Using (3) and (4), we thus get

$$
\begin{aligned}
R &= \min_{p'} H_{mon}(p') \\
&= \min_{\sigma \text{ permutation}} H_{mon}(\sigma(p)) \in [H(p), H(p)+1].
\end{aligned}
$$

It remains to prove that the optimal rate is always bounded above by $H(p)+1$. For this purpose, rearrange the vectors $|\psi_i\rangle$ in decreasing order such that $p_1 \geq p_2 \geq \cdots \geq p_n$. Let p' be the sequential projection which is constructed by getting through the list of $|\psi_i\rangle$'s in that order. As before, denote by Ψ_j the set of $|\psi_i\rangle$'s that have been collected in step j of the construction of p'. Let

$$\ell_i := \lceil -\log_2 \max_{|\psi\rangle \in \Psi_i} p(|\psi\rangle) \rceil.$$

By construction, $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_d$ and the Kraft inequality holds for the ℓ_i. Thus,

$$
\begin{aligned}
R &\leq H_{mon}(p') \leq \sum_{i=1}^{d} p'_i \ell_i \\
&= \sum_{i=1}^{d} \left(\sum_{|\psi\rangle \in \Psi_i} P(|\psi\rangle) \right) \lceil -\log_2 \max_{|\psi\rangle \in \Psi_i} p(|\psi\rangle) \rceil \\
&\leq \sum_{i=1}^{d} \sum_{|\psi\rangle \in \Psi_i} p(|\psi\rangle) \lceil -\log_2 p(|\psi\rangle) \rceil \leq H(p) + 1.
\end{aligned}
$$

This proves the statement of the theorem. ■

VII. CONCLUSIONS

We have given a method for lossless compression of unknown outputs of single quantum information sources which minimizes the code's expected base length, and we have calculated the corresponding optimal compression rate (Theorem VI 4). Moreover, we have explained a simple model of an always-open channel which admits the lossless transmission of the indeterminate-length code words and we have explained that compression can reduce transmission errors for those channels.

As our approach quantifies the rate in terms of the base length, it complements work by Schumacher and Westmoreland [7] who have given the optimal rate for average length compression. Furthermore, we have demonstrated how to apply the theory of prefix-free subspaces to quantum information. In short, prefix-free quantum strings allow sequential compression in the case of several quantum information sources by concatenating the corresponding code words. The concatenation can be accomplished physically (Theorem V 6), even in the case of prefix-free subspaces which are more general then in Schumacher and Westmoreland's sense (cf. Example V 3).

At this point, it remains open if there is a simple formula for the optimal asymptotic compression rate $\lim_{n\to\infty}\frac{1}{n}R(\mathcal{E}^{\otimes n})$ in the case of n copies of a single source $\mathcal{E}$, apart from the upper bound $R(\mathcal{E})$. Also, it would be nice to have an example of a physical situation where base length compression is better suited to reduce transmission errors for channels than average length compression (cf. Sec. IV). Even though the optimal asymptotic compression rate is not given in this paper, the result is optimal for the case of a sequence of several sources $\mathcal{E}_1\otimes\mathcal{E}_2\otimes\cdots\otimes\mathcal{E}_k$ which are not known in advance and have to be compressed sequentially.

Many open questions in quantum information theory, such as entanglement catalysis [15], are phrased as "How can this state be transformed into that state exactly and without error subject to these conditions?" Perhaps lossless quantum base length compression can be applied to some of these questions. Boström and Felbinger [4] stated that lossless quantum compression may also have applications in cryptography. Perhaps it can be used to minimize the probability that an eavesdropper discovers any information at all, rather than the average information that the eavesdropper discovers [14].

Another possible connection to existing work is in the definition of quantum Kolmogorov complexity by Berthiaume *et al.* [16]. They define the complexity of a quantum bit string as the length of its shortest determinate-length description. Therefore we might expect there to be a close correlation between this kind of complexity and the rate of compression described in this paper, in the same way that there is a close correlation between classical Kolmogorov complexity and Shannon entropy.

Apart from possible applications, one purpose of this paper was to show that prefix-free quantum bit strings are a mathematical structure with nice properties that can be useful in quantum information theory. It might be interesting to study them in more detail, in particular in connection to possible quantum versions of algorithmic probability.

ACKNOWLEDGMENTS

We thank Nihat Ay, Sougato Bose, Arleta Szkoła, Andreas Winter, and the members of the Information Theory and Cognitive Systems Group at MPI Leipzig for interesting and helpful discussions. The work of C. Rogers and R. Nagarajan was partially supported by UK EPSRC grants (No. GR/S34090 and No. EP/E00623X/1) and the EU Sixth Framework Programme (Project SecoQC: Development of a Global Network for Secure Communication based on Quantum Cryptography).

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications (Wiley, New York, 1991).
[2] B. Schumacher, Phys. Rev. A **51**, 2738 (1995).
[3] R. Ahlswede and N. Cai, IEEE Trans. Inf. Theory **50**, 1208 (2004).
[4] K. Bostroem and T. Felbinger, Phys. Rev. A **65**, 032313 (2002).
[5] S. Braunstein, C. Fuchs, D. Gottesman, and H. Lo., IEEE Trans. Inf. Theory **46**, 1644 (2000).
[6] M. Koashi and N. Imoto, Phys. Rev. Lett. **89**, 097904 (2002).
[7] B. Schumacher and M. D. Westmoreland, Phys. Rev. A **64**, 042304 (2001).
[8] M. Müller and C. Rogers in *Proceedings of the 2008 International Conference on Information Theory and Statistical Learning* (CSREA Press, Las Vegas, 2008).
[9] I. Chuang and D. Modha, IEEE Trans. Inf. Theory **46**, 1104 (2000).
[10] P. Hayden, R. Jozsa, and A. Winter, J. Math. Phys. **43**, 4404 (2002).
[11] M. Hayashi and K. Matsumoto, Phys. Rev. A **66**, 022311 (2002).
[12] L. Rallan and V. Vedral, Phys. Rev. A **68**, 042309 (2003).
[13] G. Bowen and R. Nagarajan, IEEE Trans. Inf. Theory **51**, 320 (2005).
[14] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2000).
[15] D. Jonathan and M. B. Plenio, Phys. Rev. Lett. **83**, 3566 (1999).
[16] A. Berthiaume, W. van Dam, and S. Laplante, J. Comput. Syst. Sci. **63**, 201 (2001).