

Lecture 4: Quantum typicality and quantum data compression

Mark M. Wilde*

One of the fundamental tasks in classical information theory is the compression of information. Given access to many uses of a noiseless classical channel, what is the best that a sender and receiver can make of this resource for compressed data transmission? Shannon's compression theorem states that the Shannon entropy is the fundamental limit for the compression rate in the IID setting. That is, if one compresses at a rate above the Shannon entropy, then it is possible to recover the compressed data perfectly in the asymptotic limit, and otherwise, it is not possible to do so. This theorem establishes the prominent role of the entropy in Shannon's theory of information.

In the quantum world, it very well could be that one day a sender and a receiver would have many uses of a noiseless quantum channel available,¹ and the sender could use this resource to transmit compressed quantum information. But what exactly does this mean in the quantum setting? A simple model of a quantum information source is an ensemble of quantum states $\{p_X(x), |\psi_x\rangle\}$, i.e., the source outputs the state $|\psi_x\rangle$ with probability $p_X(x)$, and the states $\{|\psi_x\rangle\}$ do not necessarily have to form an orthonormal basis. Let us suppose for the moment that the classical data x is available as well, even though this might not necessarily be the case in practice. A naive strategy for compressing this quantum information source would be to ignore the quantum states coming out, handle the classical data instead, and exploit Shannon's compression protocol. That is, the sender compresses the sequence x^n emitted from the quantum information source at a rate equal to the Shannon entropy $H(X)$, sends the compressed classical bits over the noiseless quantum channels, the receiver reproduces the classical sequence x^n at his end, and finally reconstructs the sequence $|\psi_{x^n}\rangle$ of quantum states corresponding to the classical sequence x^n .

The above strategy will certainly work, but it makes no use of the fact that the noiseless quantum channels are quantum! It is clear that noiseless quantum channels will be expensive in practice, and the above strategy is wasteful in this sense because it could have merely exploited classical channels (channels that cannot preserve superpositions) to achieve the same goals. Schumacher compression is a strategy that makes effective use of noiseless quantum channels to compress a quantum information source down to a rate equal to the von Neumann entropy. This has a great benefit from a practical standpoint—the von Neumann entropy of a quantum information source is strictly lower than the source's Shannon entropy if the states in the ensemble are non-orthogonal. In order to execute the protocol, the sender and receiver simply need to know the density operator $\rho \equiv \sum_x p_X(x) |\psi_x\rangle\langle\psi_x|$ of the source. Furthermore, Schumacher compression is provably optimal in

*Mark M. Wilde is with the Department of Physics and Astronomy and the Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803. These lecture notes are available under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. Much of the material is from the book preprint "From Classical to Quantum Shannon Theory" available as [arXiv:1106.1445](https://arxiv.org/abs/1106.1445).

¹How we hope so! If working, coherent fault-tolerant quantum computers come along one day, they stand to benefit from quantum compression protocols.

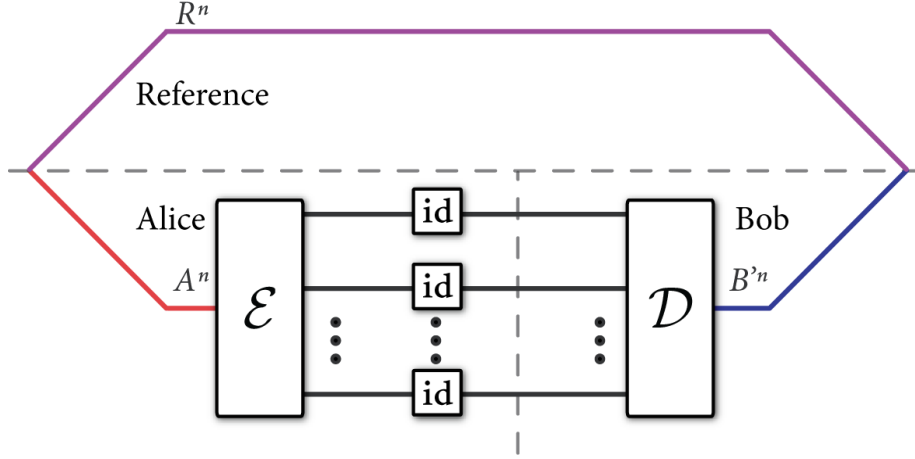


Figure 1: The most general protocol for quantum compression. Alice begins with the output of some quantum information source whose density operator is $\rho^{\otimes n}$ on some system A^n . The inaccessible reference system holds the purification of this density operator. She performs some CPTP encoding map \mathcal{E} , sends the compressed qubits through 2^{nR} uses of a noiseless quantum channel, and Bob performs some CPTP decoding map \mathcal{D} to decompress the qubits. The scheme is successful if the difference between the initial state and the final state is negligible in the asymptotic limit $n \rightarrow \infty$.

the sense that any protocol that compresses a quantum information source of the above form at a rate below the von Neumann entropy cannot have a vanishing error in the asymptotic limit.

Schumacher compression thus gives an operational interpretation of the von Neumann entropy as the fundamental limit on the rate of quantum data compression. Also, it sets the term “qubit” on a firm foundation in an information-theoretic sense as a measure of the amount of quantum information “contained” in a quantum information source.

We begin this lecture by giving the details of the general information processing task corresponding to quantum data compression. We then prove that the von Neumann entropy is an achievable rate of compression.

1 The Information Processing Task

We first overview the general task that any quantum compression protocol attempts to accomplish. Three parameters n , R , and ϵ corresponding to the length of the original quantum data sequence, the rate, and the error, respectively, characterize any such protocol. An $(n, R + \delta, \epsilon)$ quantum compression code consists of four steps: state preparation, encoding, transmission, and decoding. Figure 1 depicts a general protocol for quantum compression.

State Preparation. The quantum information source outputs a sequence $|\psi_{x^n}\rangle^{A^n}$ of quantum states according to the ensemble $\{p_X(x), |\psi_x\rangle\}$ where

$$|\psi_{x^n}\rangle^{A^n} \equiv |\psi_{x_1}\rangle^{A_1} \otimes \cdots \otimes |\psi_{x_n}\rangle^{A_n}. \quad (1)$$

The density operator, from the perspective of someone ignorant of the classical sequence x^n , is

equal to the tensor power state $\rho^{\otimes n}$ where

$$\rho \equiv \sum_x p_X(x) |\psi_x\rangle \langle \psi_x|. \quad (2)$$

Also, we can think about the purification of the above density operator. That is, an equivalent mathematical picture is to imagine that the quantum information source produces states of the form

$$|\varphi_\rho\rangle^{RA} \equiv \sum_x \sqrt{p_X(x)} |x\rangle^R |\psi_x\rangle^A, \quad (3)$$

where R is the label for an inaccessible reference system (not to be confused with the rate R !). The resulting IID state produced is $(|\varphi_\rho\rangle^{RA})^{\otimes n}$.

Encoding. Alice encodes the systems A^n according to some CPTP compression map $\mathcal{E}^{A^n \rightarrow W}$ where W is a quantum system of size 2^{nR} . Recall that R is the rate of compression:

$$R = \frac{1}{n} \log d_W - \delta, \quad (4)$$

where d_W is the dimension of system W and δ is an arbitrarily small positive number.

Transmission. Alice transmits the system W to Bob using $n(R + \delta)$ noiseless qubit channels.

Decoding. Bob sends the system W through a decompression map $\mathcal{D}^{W \rightarrow \hat{A}^n}$.

The protocol has ϵ error if the compressed and decompressed state is ϵ -close in trace distance to the original state $(|\varphi_\rho\rangle^{RA})^{\otimes n}$:

$$\left\| (\varphi_\rho^{RA})^{\otimes n} - (\mathcal{D}^{W \rightarrow \hat{A}^n} \circ \mathcal{E}^{A^n \rightarrow W})((\varphi_\rho^{RA})^{\otimes n}) \right\|_1 \leq \epsilon. \quad (5)$$

2 The Quantum Data Compression Theorem

We say that a quantum compression rate R is *achievable* if there exists an $(n, R + \delta, \epsilon)$ quantum compression code for all $\delta, \epsilon > 0$ and all sufficiently large n . Schumacher's compression theorem establishes the von Neumann entropy as the fundamental limit on quantum data compression.

Theorem 1 (Quantum Data Compression) *Suppose that ρ^A is the density operator of the quantum information source. Then the von Neumann entropy $H(A)_\rho$ is the smallest achievable rate R for quantum data compression:*

$$\inf\{R : R \text{ is achievable}\} = H(A)_\rho. \quad (6)$$

In this lecture, we will only prove the direct part of this theorem (that is, we will exhibit a coding scheme that achieves the von Neumann entropy rate of compression). The main technical tool required in proving the direct part is the notion of quantum typicality, which we describe in the next section.

2.1 Quantum Typicality

2.1.1 The Typical Subspace

Our first task is to establish the notion of a quantum information source. It is analogous to the notion of a classical information source, in the sense that the source randomly outputs a quantum state according to some probability distribution, but the states that it outputs do not necessarily have to be distinguishable as in the classical case.

Definition 2 (Quantum Information Source) *A quantum information source is some device that randomly emits pure qudit states in a Hilbert space \mathcal{H}_X of size $|\mathcal{X}|$.*

We use the symbol X to denote the quantum system for the quantum information source in addition to denoting the Hilbert space in which the state lives. Suppose that the source outputs states $|\psi_y\rangle$ randomly according to some probability distribution $p_Y(y)$. Note that the states $|\psi_y\rangle$ do not necessarily have to form an orthonormal set. Then the density operator ρ^X of the source is the expected state emitted:

$$\rho^X \equiv \mathbb{E}_Y\{|\psi_Y\rangle\langle\psi_Y|\} = \sum_y p_Y(y) |\psi_y\rangle\langle\psi_y|. \quad (7)$$

There are many decompositions of a density operator as a convex sum of rank-one projectors (and the above decomposition is one such example), but perhaps the most important decomposition is a spectral decomposition of the density operator ρ :

$$\rho^X = \sum_{x \in \mathcal{X}} p_X(x) |x\rangle\langle x|^X. \quad (8)$$

The above states $|x\rangle^X$ are eigenvectors of ρ^X and form a complete orthonormal basis for Hilbert space \mathcal{H}_X , and the non-negative, convex real numbers $p_X(x)$ are the eigenvalues of ρ^X .

We have written the states $|x\rangle^X$ and the eigenvalues $p_X(x)$ in a suggestive notation because it is actually possible to think of our quantum source as a classical information source—the emitted states $\{|x\rangle^X\}_{x \in \mathcal{X}}$ are orthonormal and each corresponding eigenvalue $p_X(x)$ acts as a probability for choosing $|x\rangle^X$. We can say that our source is classical because it is emitting the orthogonal, and thus distinguishable, states $|x\rangle^X$ with probability $p_X(x)$. This description is equivalent to the ensemble $\{p_Y(y), |\psi_y\rangle\}_y$ because the two ensembles lead to the same density operator (recall that two ensembles that have the same density operator are essentially equivalent because they lead to the same probabilities for outcomes of any measurement performed on the system). Our quantum information source then corresponds to the pure-state ensemble:

$$\{p_X(x), |x\rangle^X\}_{x \in \mathcal{X}}. \quad (9)$$

Recall that the von Neumann entropy $H(X)$ of the density operator ρ^X is as follows:

$$H(X)_\rho \equiv -\text{Tr}\{\rho^X \log \rho^X\}. \quad (10)$$

It is straightforward to show that the von Neumann entropy $H(X)_\rho$ is equal to the Shannon entropy $H(X)$ of a random variable X with distribution $p_X(x)$ because the basis states $|x\rangle^X$ are orthonormal.

Suppose now that the quantum information source emits a large number n of random quantum states so that the density operator describing the emitted state is as follows:

$$\rho^{X^n} \equiv \underbrace{\rho^{X_1} \otimes \dots \otimes \rho^{X_n}}_{n \text{ times}} = (\rho^X)^{\otimes n}. \quad (11)$$

The labels X_1, \dots, X_n denote the Hilbert spaces in which the different quantum systems live, but the density operator is the same for each Hilbert space X_1, \dots, X_n and is equal to ρ^X . The above

description of a quantum source is within the independent and identically distributed (IID) setting for the quantum domain. The spectral decomposition of the state in (11) is as follows:

$$\rho^{X^n} = \sum_{x_1 \in \mathcal{X}} p_X(x_1) |x_1\rangle\langle x_1|^{X_1} \otimes \cdots \otimes \sum_{x_n \in \mathcal{X}} p_X(x_n) |x_n\rangle\langle x_n|^{X_n} \quad (12)$$

$$= \sum_{x_1, \dots, x_n \in \mathcal{X}} p_X(x_1) \cdots p_X(x_n) |x_1\rangle \cdots |x_n\rangle\langle x_1| \cdots \langle x_n|^{X_1, \dots, X_n} \quad (13)$$

$$= \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) |x^n\rangle\langle x^n|^{X^n}, \quad (14)$$

where we employ the shorthand:

$$p_{X^n}(x^n) \equiv p_X(x_1) \cdots p_X(x_n), \quad |x^n\rangle^{X^n} \equiv |x_1\rangle^{X_1} \cdots |x_n\rangle^{X_n}. \quad (15)$$

The above quantum description of the density operator is essentially equivalent to the classical picture of n realizations of random variable X with each eigenvalue $p_{X_1}(x_1) \cdots p_{X_n}(x_n)$ acting as a probability because the set of states $\{|x_1\rangle \cdots |x_n\rangle^{X_1, \dots, X_n}\}_{x_1, \dots, x_n \in \mathcal{X}}$ is an orthonormal set.

Definition 3 (Typical Subspace) *The δ -typical subspace $T_\delta^{X^n}$ is a subspace of the full Hilbert space X_1, \dots, X_n and is associated with many copies of a density operator, such as the one in (8). It is spanned by states $|x^n\rangle^{X^n}$ whose corresponding classical sequences x^n are δ -typical:*

$$T_\delta^{X^n} \equiv \text{span}\left\{|x^n\rangle^{X^n} : |-\log(p_{X^n}(x^n))/n - H(X)| \leq \delta\right\}, \quad (16)$$

where $p_X(x)$ is the distribution from the spectral decomposition of ρ in (8). We could also denote the typical subspace as $T_{\rho, \delta}^{X^n}$ if we would like to make the dependence of the space on ρ more explicit.

2.1.2 The Typical Subspace Measurement

The definition of the typical subspace (Definition 3) gives a way to divide up the Hilbert space of n qudits into two subspaces: the typical subspace and the atypical subspace. The properties of the typical subspace are similar to what we found for the properties of typical sequences. That is, the typical subspace is exponentially smaller than the full Hilbert space of n qudits, yet it contains nearly all of the probability (in a sense that we show below). The intuition for these properties of the typical subspace is the same as it is classically once we have a spectral decomposition of a density operator.

The *typical projector* is a projector onto the typical subspace, and the complementary projector projects onto the atypical subspace. These projectors play an important operational role in quantum Shannon theory because we can construct a quantum measurement from them. That is, this measurement is the best way of asking the question, “Is the state typical or not?” because it minimally disturbs the state while still retrieving this one bit of information.

Definition 4 (Typical Projector) *Let $\Pi_\delta^{X^n}$ denote the typical projector for the typical subspace of a density operator ρ^X with spectral decomposition in (8). It is a projector onto the typical subspace:*

$$\Pi_\delta^{X^n} \equiv \sum_{x^n \in T_\delta^{X^n}} |x^n\rangle\langle x^n|^{X^n}, \quad (17)$$

where it is implicit that the x^n below the summation is a classical sequence in the typical set $T_\delta^{X^n}$, and the state $|x^n\rangle$ is a quantum state given in (15) and associated with the classical sequence x^n via the spectral decomposition of ρ in (8). We can also denote the typical projector as $\Pi_{\rho,\delta}^{X^n}$ if we would like to make its dependence on ρ explicit.

The action of multiplying the density operator ρ^{X^n} by the typical projector $\Pi_\delta^{X^n}$ is to select out all the basis states of ρ^{X^n} that are in the typical subspace and form a “sliced” operator $\tilde{\rho}^{X^n}$ that is close to the original density operator ρ^{X^n} :

$$\tilde{\rho}^{X^n} \equiv \Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) |x^n\rangle \langle x^n|. \quad (18)$$

That is, the effect of projecting a state onto the typical subspace $T_\delta^{X^n}$ is to “slice” out any component of the state ρ^{X^n} that does not lie in the typical subspace $T_\delta^{X^n}$.

Exercise 5 Show that the typical projector $\Pi_\delta^{X^n}$ commutes with the density operator ρ^{X^n} :

$$\rho^{X^n} \Pi_\delta^{X^n} = \Pi_\delta^{X^n} \rho^{X^n}. \quad (19)$$

The typical projector allows us to formulate an operational method for delicately asking the question: “Is the state typical or not?” We can construct a quantum measurement that consists of two outcomes: the outcome “1” reveals that the state is in the typical subspace, and “0” reveals that it is not. This typical subspace measurement is often one of the first important steps in most protocols in quantum Shannon theory.

Definition 6 (Typical Subspace Measurement) The following map is a quantum measurement map that realizes the typical subspace measurement:

$$\sigma \rightarrow (I - \Pi_\delta^{X^n}) \sigma (I - \Pi_\delta^{X^n}) \otimes |0\rangle \langle 0| + \Pi_\delta^{X^n} \sigma \Pi_\delta^{X^n} \otimes |1\rangle \langle 1|, \quad (20)$$

where σ is some quantum state living in the Hilbert space X^n . It associates a classical register with the outcome of the measurement—the value of the classical register is $|0\rangle$ for the support of the state σ that is not in the typical subspace, and it is equal to $|1\rangle$ for the support of the state σ that is in the typical subspace.

2.1.3 Properties of the Typical Subspace

The typical subspace $T_\delta^{X^n}$ enjoys several useful properties that are “quantized” versions of the typical sequence properties:

Property 7 (Unit Probability) Suppose that we perform a typical subspace measurement of a state ρ^{X^n} . Then the probability that the quantum state ρ^{X^n} is in the typical subspace $T_\delta^{X^n}$ approaches one as n becomes large:

$$\forall \epsilon > 0 \quad \text{Tr}\{\Pi_\delta^{X^n} \rho^{X^n}\} \geq 1 - \epsilon \quad \text{for sufficiently large } n, \quad (21)$$

where $\Pi_\delta^{X^n}$ is the typical subspace projector from Definition 4.

Property 8 (Exponentially Small Dimension) *The dimension $\dim(T_\delta^{X^n})$ of the δ -typical subspace is exponentially smaller than the dimension $|\mathcal{X}|^n$ of the entire space of quantum states when the output of the quantum information source is not maximally mixed. We formally state this property as follows:*

$$\text{Tr}\{\Pi_\delta^{X^n}\} \leq 2^{n(H(X)+c\delta)}, \quad (22)$$

where c is some constant that depends on whether we employ the weak or strong notion of typicality. We can also lower bound the dimension $\dim(T_\delta^{X^n})$ of the δ -typical subspace when n is sufficiently large:

$$\forall \epsilon > 0 \quad \text{Tr}\{\Pi_\delta^{X^n}\} \geq (1 - \epsilon)2^{n(H(X)-c\delta)} \quad \text{for sufficiently large } n. \quad (23)$$

Property 9 (Equipartition) *The operator $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ corresponds to a “slicing” of the density operator ρ^{X^n} where we slice out and keep only the part with support in the typical subspace. We can then bound all of the eigenvalues of the sliced operator $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ as follows:*

$$2^{-n(H(X)+c\delta)} \Pi_\delta^{X^n} \leq \Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n} \leq 2^{-n(H(X)-c\delta)} \Pi_\delta^{X^n}. \quad (24)$$

The above inequality is an operator inequality. It is a statement about the eigenvalues of the operators $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ and $\Pi_\delta^{X^n}$, and these operators have the same eigenvectors because they commute. Therefore, the above inequality is equivalent to the following inequality that applies in the classical case:

$$\forall x^n \in T_\delta^{X^n} : 2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}. \quad (25)$$

This equivalence holds because each probability $p_{X^n}(x^n)$ is an eigenvalue of $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$.

The dimension $\dim(T_\delta^{X^n})$ of the δ -typical subspace is approximately equal to the dimension $|\mathcal{X}|^n$ of the entire space only when the density operator of the quantum information source is maximally mixed because

$$\text{Tr}\{\Pi_\delta^{X^n}\} \leq |\mathcal{X}|^n \cdot 2^{n\delta} \simeq |\mathcal{X}|^n. \quad (26)$$

The proofs of the above properties are essentially identical to those from the theory of classical typicality (see arXiv:1106.1445). We leave the proofs as the three exercises below.

Exercise 10 *Prove the Unit Probability Property of the δ -typical subspace (Property 7). First show that the probability that many copies of a density operator is in the δ -typical subspace is equal to the probability that a random sequence is δ -typical:*

$$\text{Tr}\{\Pi_\delta^{X^n} \rho^{X^n}\} = \Pr\{X^n \in T_\delta^{X^n}\}. \quad (27)$$

Exercise 11 *Prove the Exponentially Small Dimension Property of the δ -typical subspace (Property 8). First show that the trace of the typical projector $\Pi_\delta^{X^n}$ is equal to the dimension of the typical subspace $T_\delta^{X^n}$:*

$$\dim(T_\delta^{X^n}) = \text{Tr}\{\Pi_\delta^{X^n}\}. \quad (28)$$

Then prove the property.

Exercise 12 *Prove the Equipartition Property of the δ -typical subspace (Property 9). First show that*

$$\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) |x^n\rangle \langle x^n|^{X^n}, \quad (29)$$

and then argue the proof.

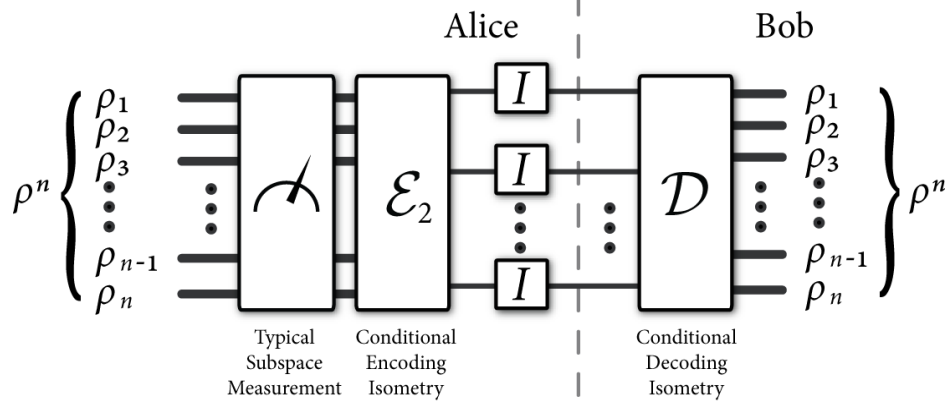


Figure 2: Schumacher's compression protocol. Alice begins with many copies of the output of the quantum information source. She performs a measurement onto the typical subspace corresponding to the state ρ and then performs a compression isometry of the typical subspace to a space of size $2^{n[H(\rho)+\delta]}$ qubits. She transmits these compressed qubits over $n[H(\rho) + \delta]$ uses of a noiseless quantum channel. Bob performs the inverse of the isometry to uncompress the qubits. The protocol is successful in the asymptotic limit due to the properties of typical subspaces.

Exercise 13 Show that the purity $\text{Tr}\{(\tilde{\rho}^{X^n})^2\}$ of the sliced state $\tilde{\rho}^{X^n}$ satisfies the following bound for sufficiently large n and any $\epsilon > 0$ (use weak quantum typicality):

$$(1 - \epsilon)2^{-n(H(X)+\delta)} \leq \text{Tr}\{(\tilde{\rho}^{X^n})^2\} \leq 2^{-n(H(X)-\delta)}. \quad (30)$$

Exercise 14 Show that the following bounds hold for the zero-norm and the ∞ -norm of the sliced state $\tilde{\rho}^{X^n}$ for any $\epsilon > 0$ and sufficiently large n :

$$(1 - \epsilon)2^{n(H(X)-\delta)} \leq \|\tilde{\rho}^{X^n}\|_0 \leq 2^{n(H(X)+\delta)}, \quad (31)$$

$$2^{-n(H(X)+\delta)} \leq \|\tilde{\rho}^{X^n}\|_\infty \leq 2^{-n(H(X)-\delta)}. \quad (32)$$

(Recall that the zero-norm of an operator is equal to the size of its support and that the infinity norm is equal to its maximum eigenvalue. Again use weak quantum typicality.)

2.2 The Direct Coding Theorem

Schumacher's compression protocol demonstrates that the von Neumann entropy $H(A)_\rho$ is an achievable rate for quantum data compression. It is remarkably similar to Shannon's compression protocol, but it has some subtle differences that are necessary for the quantum setting. The basic steps of the encoding are to perform a typical subspace measurement and an isometry that compresses the typical subspace. The decoder then performs the inverse of the isometry to decompress the state. The protocol is successful if the typical subspace measurement successfully projects onto the typical subspace, and it fails otherwise. Just like in the classical case, the law of large numbers guarantees that the protocol is successful in the asymptotic limit as $n \rightarrow \infty$. Figure 2 provides an illustration of the protocol, and we now provide a rigorous argument.

Alice begins with many copies of the state $(\varphi_\rho^{RA})^{\otimes n}$. Suppose that the spectral decomposition of ρ is as follows:

$$\rho = \sum_z p_Z(z) |z\rangle\langle z|, \quad (33)$$

where $p_Z(z)$ is some probability distribution, and $\{|z\rangle\}$ is some orthonormal basis. Her first step $\mathcal{E}_1^{A^n \rightarrow Y A^n}$ is to perform a typical subspace measurement of the form in (6) onto the typical subspace of A^n , where the typical projector is with respect to the density operator ρ . The action of $\mathcal{E}_1^{A^n \rightarrow Y A^n}$ on a general state σ^{A^n} is

$$\mathcal{E}_1^{A^n \rightarrow Y A^n}(\sigma^{A^n}) \equiv |0\rangle\langle 0|^Y \otimes (I - \Pi_\delta^{A^n}) \sigma^{A^n} (I - \Pi_\delta^{A^n}) + |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n} \sigma^{A^n} \Pi_\delta^{A^n}, \quad (34)$$

and the classically-correlated flag bit Y indicates whether the typical subspace projection $\Pi_\delta^{A^n}$ is successful or unsuccessful. In the Shannon compression protocol, we exploit a one-to-one function f that maps from the set of typical sequences to a set of binary sequences $\{0, 1\}^{n[H(\rho)+\delta]}$. Now, we can construct an isometry U_f that is a coherent version of this classical function f . It simply maps the orthonormal basis $\{|z^n\rangle^{A^n}\}$ to the basis $\{|f(z^n)\rangle^W\}$:

$$U_f \equiv \sum_{z^n \in T_\delta^{Z^n}} |f(z^n)\rangle^W \langle z^n|^{A^n}, \quad (35)$$

where Z is a random variable corresponding to the distribution $p_Z(z)$ so that $T_\delta^{Z^n}$ is its typical set. The above operator is an isometry because the input space is a subspace of size at most $2^{n[H(\rho)+\delta]}$ (recall Property 8) embedded in a larger space of size 2^n (at least for qubits) and the output space is of size at most $2^{n[H(\rho)+\delta]}$. So her next step $\mathcal{E}_2^{Y A^n \rightarrow Y W}$ is to perform the isometric compression conditional on the flag bit Y being equal to one, and the action of $\mathcal{E}_2^{Y A^n \rightarrow Y W}$ on a general classical-quantum state $\sigma^{Y A^n} \equiv |0\rangle\langle 0|^Y \otimes \sigma_0^{A^n} + |1\rangle\langle 1|^Y \otimes \sigma_1^{A^n}$ is as follows:

$$\mathcal{E}_2^{Y A^n \rightarrow Y W}(\sigma^{Y A^n}) \equiv |0\rangle\langle 0|^Y \otimes \text{Tr}\{\sigma_0^{A^n}\} |e\rangle\langle e|^W + |1\rangle\langle 1|^Y \otimes U_f \sigma_1^{A^n} U_f^\dagger, \quad (36)$$

where $|e\rangle^W$ is some error flag orthogonal to all of the states $\{|f(\phi_{x^n})\rangle^W\}_{\phi_{x^n} \in T_\delta^{Z^n}}$. This last step completes the details of her encoder $\mathcal{E}^{A^n \rightarrow Y W}$, and the action of it on the initial state is

$$\mathcal{E}^{A^n \rightarrow Y W}((\varphi_\rho^{RA})^{\otimes n}) \equiv (\mathcal{E}_2^{Y A^n \rightarrow Y W} \circ \mathcal{E}_1^{A^n \rightarrow Y A^n})((\varphi_\rho^{RA})^{\otimes n}). \quad (37)$$

Alice then transmits all of the compressed qubits over $n[H(\rho) + \delta] + 1$ uses of the noiseless qubit channel.

Bob's decoding $\mathcal{D}^{Y W \rightarrow A^n}$ performs the inverse of the isometry conditional on the flag bit being equal to one and otherwise maps to some other state $|e\rangle^{A^n}$ outside of the typical subspace. The action of the decoder on some general classical-quantum state $\sigma^{Y W} \equiv |0\rangle\langle 0|^Y \otimes \sigma_0^W + |1\rangle\langle 1|^Y \otimes \sigma_1^W$ is

$$\mathcal{D}_1^{Y W \rightarrow Y A^n}(\sigma^{Y W}) \equiv |0\rangle\langle 0|^Y \otimes \text{Tr}\{\sigma_0^W\} |e\rangle\langle e|^{A^n} + |1\rangle\langle 1|^Y \otimes U_f^\dagger \sigma_1^W U_f. \quad (38)$$

The final part of the decoder is to discard the classical flag bit: $\mathcal{D}_2^{Y A^n \rightarrow A^n} \equiv \text{Tr}_Y\{\cdot\}$. Then $\mathcal{D}^{Y W \rightarrow A^n} \equiv \mathcal{D}_2^{Y A^n \rightarrow A^n} \circ \mathcal{D}_1^{Y W \rightarrow Y A^n}$.

We now can analyze how this protocol performs with respect to our performance criterion in (5). Consider the following chain of inequalities:

$$\begin{aligned} & \left\| (\varphi_\rho^{RA})^{\otimes n} - (\mathcal{D}^{YW \rightarrow A^n} \circ \mathcal{E}^{A^n \rightarrow YW})((\varphi_\rho^{RA})^{\otimes n}) \right\|_1 \\ &= \left\| \text{Tr}_Y \left\{ |1\rangle\langle 1|^Y \otimes (\varphi_\rho^{RA})^{\otimes n} \right\} - (\mathcal{D}^{YW \rightarrow A^n} \circ \mathcal{E}^{A^n \rightarrow YW})((\varphi_\rho^{RA})^{\otimes n}) \right\|_1 \end{aligned} \quad (39)$$

$$\leq \left\| |1\rangle\langle 1|^Y \otimes (\varphi_\rho^{RA})^{\otimes n} - (\mathcal{D}_1^{YW \rightarrow A^n} \circ \mathcal{E}^{A^n \rightarrow YW})((\varphi_\rho^{RA})^{\otimes n}) \right\|_1 \quad (40)$$

$$= \left\| \begin{pmatrix} |1\rangle\langle 1|^Y \otimes (\varphi_\rho^{RA})^{\otimes n} - \\ |0\rangle\langle 0|^Y \otimes \text{Tr} \left\{ (I - \Pi_\delta^{A^n}) (\varphi_\rho^{RA})^{\otimes n} \right\} |e\rangle\langle e|^{A^n} \\ + |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n} (\varphi_\rho^{RA})^{\otimes n} \Pi_\delta^{A^n} \end{pmatrix} \right\|_1 \quad (41)$$

The first equality follows by adding a flag bit $|1\rangle^Y$ to $(\varphi_\rho^{RA})^{\otimes n}$ and tracing it out. The first inequality follows from monotonicity of trace distance under the discarding of subsystems. The second equality follows by evaluating the map $\mathcal{D}_1^{YW \rightarrow A^n} \circ \mathcal{E}^{A^n \rightarrow YW}$ on the state $(\varphi_\rho^{RA})^{\otimes n}$. Continuing, we have

$$\begin{aligned} & \leq \left\| |1\rangle\langle 1|^Y \otimes (\varphi_\rho^{RA})^{\otimes n} - |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n} (\varphi_\rho^{RA})^{\otimes n} \Pi_\delta^{A^n} \right\|_1 \\ & \quad + \left\| |0\rangle\langle 0|^Y \otimes \text{Tr} \left\{ (I - \Pi_\delta^{A^n}) (\varphi_\rho^{RA})^{\otimes n} \right\} |e\rangle\langle e|^{A^n} \right\|_1 \end{aligned} \quad (42)$$

$$= \left\| (\varphi_\rho^{RA})^{\otimes n} - \Pi_\delta^{A^n} (\varphi_\rho^{RA})^{\otimes n} \Pi_\delta^{A^n} \right\|_1 + \text{Tr} \left\{ (I - \Pi_\delta^{A^n}) (\varphi_\rho^{RA})^{\otimes n} \right\} \quad (43)$$

$$\leq 2\sqrt{\epsilon} + \epsilon \quad (44)$$

The first inequality follows from the triangle inequality for trace distance. The equality uses the facts $\|\rho \otimes \sigma - \omega \otimes \sigma\|_1 = \|\rho - \omega\|_1 \|\sigma\|_1 = \|\rho - \omega\|_1$ and $\|b\rho\|_1 = |b| \|\rho\|_1$ for some density operators ρ, σ , and ω and a constant b . The final inequality follows from the first property of typical subspaces:

$$\text{Tr} \left\{ \Pi_\delta^{A^n} (\varphi_\rho^{RA})^{\otimes n} \right\} = \text{Tr} \left\{ \Pi_\delta^{A^n} \rho^{\otimes n} \right\} \geq 1 - \epsilon, \quad (45)$$

and the following Gentle Operator Lemma:

Lemma 15 (Gentle Operator Lemma) *Suppose that an operator Λ such that $I \geq \Lambda \geq 0$ succeeds with high probability on the state ρ :*

$$\text{Tr} \{ \Lambda \rho \} \geq 1 - \epsilon.$$

Then the subnormalized state $\sqrt{\Lambda} \rho \sqrt{\Lambda}$ is close in trace distance to the original state ρ :

$$\left\| \sqrt{\Lambda} \rho \sqrt{\Lambda} - \rho \right\|_1 \leq 2\sqrt{\epsilon}.$$

We remark that it is important for the typical subspace measurement in (34) to be implemented as a coherent quantum measurement. That is, the only information that this measurement should learn is whether the state is typical or not. Otherwise, there would be too much disturbance to the quantum information, and the protocol would fail at the desired task of compression. Such precise control on so many qubits is possible in principle, but it is of course rather daunting to implement in practice!

3 Quantum Compression Example

We now highlight a particular example where Schumacher compression gives a big savings in compression rates if noiseless quantum channels are available. Suppose that the ensemble is of the following form:

$$\left\{ \left(\frac{1}{2}, |0\rangle \right), \left(\frac{1}{2}, |+\rangle \right) \right\}. \quad (46)$$

This ensemble is known as the Bennett-92 ensemble because it is useful in Bennett's protocol for quantum key distribution. The naive strategy would be for Alice and Bob to exploit Shannon's compression protocol. That is, Alice would ignore the quantum nature of the states, and supposing that the classical label for them were available, she would encode the classical label. Though, the entropy of the uniform distribution on two states is equal to one bit, and she would have to transmit classical messages at a rate of one bit per channel use.

A far wiser strategy is to employ Schumacher compression. The density operator of the above ensemble is

$$\frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|+\rangle\langle +|, \quad (47)$$

which has the following spectral decomposition:

$$\cos^2(\pi/8)|+\rangle\langle +| + \sin^2(\pi/8)|-\rangle\langle -|, \quad (48)$$

where

$$|+\rangle \equiv \cos(\pi/8)|0\rangle + \sin(\pi/8)|1\rangle, \quad (49)$$

$$|-\rangle \equiv \sin(\pi/8)|0\rangle - \cos(\pi/8)|1\rangle. \quad (50)$$

The binary entropy $H_2(\cos^2(\pi/8))$ of the distribution $[\cos^2(\pi/8), \sin^2(\pi/8)]$ is approximately equal to

$$0.6009 \text{ qubits}, \quad (51)$$

and thus they can save a significant amount in terms of compression rate by employing Schumacher compression. This type of savings will always occur whenever the ensemble includes non-orthogonal quantum states.

Exercise 16 *In the above example, suppose that Alice associates a classical label with the states, so that the ensemble instead is*

$$\left\{ \left(\frac{1}{2}, |0\rangle\langle 0| \otimes |0\rangle\langle 0| \right), \left(\frac{1}{2}, |1\rangle\langle 1| \otimes |+\rangle\langle +| \right) \right\}. \quad (52)$$

Does this help in reducing the amount of qubits she has to transmit to Bob?

4 Concluding Remarks

Schumacher compression was the first quantum Shannon-theoretic result discovered and is the simplest one that we encounter in this book. The proof is remarkably similar to the proof of Shannon's noiseless coding theorem, with the main difference being that we should be more careful in the

quantum case not to be learning any more information than necessary when performing measurements. The intuition that we gain for future quantum protocols is that it often suffices to consider only what happens to a high probability subspace rather than the whole space itself if our primary goal is to have a small probability of error in a communication task. In fact, this intuition is the same needed for understanding information processing tasks such as entanglement concentration, classical communication, private classical communication, and quantum communication.