# MIT 8.S372
# Quantum Information Science III
# Fall 2020

Taught by Aram Harrow, written by the class, compiled by Michael DeMarco

*Department of Physics*
*Massachusetts Institute of Technology*
aram@mit.edu, demarco@mit.edu

# Lecture topics

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 1: Sep 1, 2020

*Lecturer: Aram Harrow* *Scribe: Michael DeMarco*

## 1.1 Entanglement and Density Matrices

In quantum mechanics (QM), a *pure state* $|\psi\rangle$ is a vector in $\mathbb{C}^2, \mathbb{C}^d, \mathbb{C}^2 \otimes \mathbb{C}^2$, etc, that we use to describe a system whose state is known. On the other hand, a *mixed state* is when a system is a statistical mixture of pure states, and must be described by a *density matrix*:

$$\rho = \sum_i p_i \, |\psi_i\rangle \langle\psi_i| \in H(\mathbb{C}^d) \tag{1.1}$$

Here the system is in state $|\psi_i\rangle$ with probability $p_i$. Note that any pure state $|\psi\rangle$ has a density matrix representation as $|\psi\rangle\langle\psi|$. We shall use the notation $\psi \equiv |\psi\rangle\langle\psi|$.

As a general rule, density matrices are Hermitian matrices such that $\text{Tr }\rho = 1$ and all eigenvalues are nonnegative (often written as $\rho >= 0$). One should think of this as the quantum analog of the probability simplex, and indeed there are several notions of a probability distribution encoded in a density matrix $\rho$. If we measure a system in the natural bases $(|1\rangle ... |d\rangle)$, then $\rho_{ii}$ is the probability to find the system in the state $|i\rangle$. So we may think of the diagonal entries of $\rho$ as a probability distribution. This holds true even if we change basis, and so the eigenvalues of $\rho$ are again a probability distribution.

Mixed states can be obtained from entangled states by discarding information about a subsystem. Let our system partition into $A$ and $B$ subsystems. Then $\psi_A \equiv \text{Tr }_B\psi$. Specifically, suppose that we can write a pure state $|\psi\rangle$ as:

$$|\psi\rangle = \sum_{ij} c_{ij} \, |i\rangle \otimes |j\rangle \tag{1.2}$$

(we will sometimes omit tensor product symbols below). Then we can write the density matrix as:

$$\psi = \sum_{ijkl} c_{ij} c_{kl}^* \, |i\rangle \otimes |j\rangle \langle k| \otimes \langle l| = \sum_{ijkl} c_{ij} c_{kl}^* \, |i\rangle \langle k| \otimes |j\rangle \langle l| \tag{1.3}$$

Now, note that we may consider $\text{Tr } : L(\mathbb{C}^d) \to \mathbb{C}$, and $I : L(\mathbb{C}^d) \to \mathbb{C}^d$. So that $\text{Tr }_B = \text{Tr } \otimes I$. Accordingly, taking the trace over the B subsystem above replaces

Tr $(|j\rangle\langle l|) = \delta jl$, and so:

$$\psi_A = \text{Tr}\ _B\psi = \sum_{ijk} c_{ij}c^*_{kj}\,|i\rangle\,|j\rangle \tag{1.4}$$

If we consider $c_{ij}$ to be the entries of a (not necessarily square) matrix $C$, then we can write this as:

$$\psi_A = CC^\dagger \tag{1.5}$$

**Examples:**

1. Suppose that $C$ is rank 1, or equivalently that $c_{ij} = \alpha_i\beta^*_j$. Then $\psi$ is an unentangled product state, and $\psi_A = \alpha_i\alpha^*_j = |\alpha\rangle\langle\beta|$. Later we will see that $\psi$ is a product state $\leftrightarrow$ $\psi_A$ is pure state $\leftrightarrow$ $\psi_B$ is a pure state.

2. Suppose that $C$ has the form $C = \frac{1}{d}U$, where $U$ is a unitary matrix, and $d$ is the dimension of the matrix (necessary so that Tr $\psi = 1$). We can write $U = \sum_i |u_i\rangle\langle u_i|$, where $u_i$ are the orthonormal eigenvectors of $U$. Then we can write the quantum state as:

$$|\psi\rangle = \frac{1}{\sqrt{d}}\sum_i |u_i\rangle\,|i\rangle \tag{1.6}$$

Then one can check that

$$\psi_A = \frac{1}{d}\sum_i |u_i\rangle\langle u_i| \tag{1.7}$$

These two examples display the range of information that can be lost when we throw out a subsystem. In the first example, the $A$ subsystem remains in a pure quantum state, despite the loss of $B$. On the other hand, discarding the $B$ system destroys all correlation in $A$ in the second example, leaving $A$ in a "fully mixed state."

These phenomena are related to the *singular value decomposition* (SVD) of the matrix $C$. Let $C = UDV^\dagger$, with $U, V$ unitary and $D$ diagonal. Note that $\psi_A = CC^\dagger = UD^2U^dagger$. this implies that the eigenvalues of $\psi_A$ are the squares of the singular values of $C$ (eig$(A)$ = svd$(C)^2$). **Exercise:** Show that eig$(\psi_A)$ = eig$(\psi_B)$. This also implies that $\psi_A$ does not depend on $V$. **Exercise:** Show that $\psi_A$ is independent of unitary transformations on the B subsystem, and vice-versa.

## 1.2 Purifications (to be continued)

. The basic idea of a purification is to construct a pure state from a density matrix. Given some $\rho$ on a system $A$, can we add some subsystem $B$ and create a state $|\psi\rangle$ on systems $A$ and $B$ so that $\psi_A = \text{Tr}\,_B\psi = \rho$? In the next class, we will show that this is always possible, but not unique. This will lead to interesting results regarding bit commitment.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 2: Sep 3, 2020

*Lecturer: Aram Harrow* *Scribe: Andrey Boris Khesin, Michael DeMarco*

Today

- unitary freedom of purifications
- applications to bit commitment
- norms, trace distance and fidelity

We can use matrix considerations from the previous lecture to see how to purify quantum states. Recall that we can write:

$$|\psi\rangle = \sum_{i,j} c_{i,j} |i\rangle \otimes |j\rangle \equiv vec(C) \tag{2.1}$$

$$\psi_A = CC^\dagger \tag{2.2}$$

$$C = UDV^\dagger \implies \psi_A = UD^2U^\dagger \tag{2.3}$$

Now, given $\psi_A$, $\exists U, D$ s.t. $\psi_A = UD^2U^\dagger$. We can choose $C = UD$ or $C = UDV^\dagger$ for any unitary $V$. This allows to purify any density matrix.

Besides the choice of $V$, there is another redundancy inherent in this formalism. Suppose that we have two matrices (of the same size) $C, \tilde{C}$ such that $CC^\dagger = \tilde{C}\tilde{C}^\dagger$, with $C = UDV^\dagger$ and $\tilde{C} = \tilde{U}\tilde{D}\tilde{V}^\dagger$. Then $UD^2U^\dagger = \tilde{U}\tilde{D}^2\tilde{U}^\dagger \implies D = \tilde{D}$, with $D = \mathrm{diag}(\lambda_1\ \lambda_1\ \lambda_1\ \lambda_2\ \lambda_2\ \lambda_3)$. We have the freedom to right multiply $U$ by any unitary matrix that commutes with $D$, i.e. which acts block-diagonally on the degenerate eigenvalues in $D$:

$$CC^\dagger = \tilde{C}\tilde{C}^\dagger UD^2U^\dagger = \tilde{U}\tilde{D}^2\tilde{U}^\dagger \implies D = \tilde{D}U = \tilde{U}R \text{ for some } R \text{ such that } [R, D] = 0 \tag{2.4}$$

These considerations allow us to prove:

**Theorem 1** *Given states $|\psi\rangle_{AB}$ and $|\gamma\rangle_{AB}$, $\psi_A = \gamma_A \iff$ there exists a unitary $W$ s.t. $(I \otimes W)|\psi\rangle = |\gamma\rangle$.*

Proof: $\Longleftarrow$ is easy. $\Longrightarrow$ : Let $|\psi\rangle = vec(X)$, $|\gamma\rangle = vec(Y) = \sum_{i,j} Y_{i,j} |i\rangle \otimes |j\rangle$, with $XX^\dagger = YY^\dagger$, $X = U_1 D_1 V_1^\dagger$, $Y = U_2 D_2 V_2^\dagger$ Now, $U_1 D_1^2 U_1^\dagger = U_2 D_2^2 U_2^\dagger \implies D_1 = D_2$ Hence $U_2 = U_1 R$, for some $R$ s.t. $[R, D_1] = 0$.

On the other hand, this implies that:

$$(I \otimes W) |\psi\rangle = (I \otimes W) \sum_{i,j} X_{i,j} |i\rangle \otimes |j\rangle = \tag{2.5}$$

$$\sum_{i,j} X_{i,j} |i\rangle \otimes W |j\rangle = \tag{2.6}$$

$$\sum_{i,j,k} X_{i,j} W_{j,k} |i\rangle \otimes |k\rangle = vec(XW) \tag{2.7}$$

(Note that applying unitary to the 1st system is represented by left-multiplication on $X$) Now, since $W = V_1 R V_2^\dagger$, $XW = (U_1 D_1 V_1^\dagger)(V_1 R V_2^\dagger) = U_2 D_2 V_2^\dagger = Y$ and so the proof is complete.

Corollary: Consider two states $|\psi\rangle_{AB}$, $|\gamma\rangle_{AB'}$. Then $\psi_A = \gamma_A \iff$ either there is an isometry $V : B \to B'$ or $V : B' \to B$ that relates them. (This allows us to relax the condition that $C, \tilde{C}$ be of the same size.

**Quantum Key Distribution** We now turn to the BB84 cryptosystem as applied to quantum key distribution (QKD). Alice chooses a random bit $r$, and a random basis $b \in \{X, Z\}$

($Z$ basis: $|0\rangle$, $|1\rangle$ $X$ basis: $|+\rangle$, $|-\rangle = \frac{|0\rangle \pm |1\rangle}{\sqrt{2}}$)

Alice sends this to Bob, and he measures in a random basis $m \in \{X, Z\}$. Bob tells Alice he's measured, then both reveal bases, discard if $b! = m$, otherwise keep answer. Repeat $N$ times, get about $\frac{N}{2}$ bits. If Eve measures, she will introduce errors, which Alice and Bob can detect.At the end of this, Alice and Bob have a "key", a shared secret random string.

### Coin Flipping

In quantum mechanics, strong coin-flipping is impossible, weak coin-flipping with any bias $\epsilon > 0$ is possible. Alice can choose any bias between 0 and $\frac{1}{2} + \epsilon$, Bob can choose anything in $[\frac{1}{2} - \epsilon, 1]$.

### Bit Commitment

In bit commitment, Alice and Bob do not trust each other, so we want to devise a way for Alice to commit a bit to Bob, and we would like to store this bit for a while, without Bob being able to see it nor Alice being able to change it, before revealing it to Bob. Formally, there are two phases: Commit and Reveal. In the commit phase, Alice commits to a bit $b$, and then in the reveal phase Bob learns $b$.

A secure bit commitment most satisfy:

1. Valid: If both players are honest, Bob learns $b$ and doesn't abort
2. Hiding: After Commit phase, Bob can't learn $b$
3. Binding: During Reveal phase, Alice can convince Bob to accept only one value of $b$.

Related to bit commitment is oblivious transfer (OT): stronger than Bit Commitment, equivalent to secure multi-party computation. Alice chooses bits $(x_0, x_1)$. Bob inputs $b$, learns $x_b$. Alice learns nothing.

Generically, OT > BC > strong coin flipping.

Many of these things are possible quantumly with computational assumptions, but impossible without them. (See Urmila Mahadev + others... LWE=learning with errors)

**Theorem 2** *Information-theoretically secure quantum bit commitment is impossible.*

**Detour: Quantum Channels**

QCs encode Noisy quantum operation $\rho_A \to N(\rho_B)$ Which $N$ are allowed? (Analogous to unitary or stochastic matrices for pure-state quantum mechanics or probability)? We allow the following operations:

1. TPCP maps = Trace-preserving, completely positive linear maps
2. Kraus decomposition. $N(\rho) = \sum_k E_k \rho E_k^\dagger$, where $\{E_k\}$ are Kraus operators
3. $N(\rho) = \mathrm{tr}_E[V\rho V^\dagger]$, where $V : A \to B \otimes E$ is an isometry.

For our purposes, we will take the third option.

Now a secure bitcommitment protocol looks like Alice and bob exchanging bits, a commit phase, more exchanges, and then a reveal phase:

Alice → Bob

$\leftarrow$

$\rightarrow$

$\leftarrow$

Commit phase: state is $\rho_0$ or $\rho_1$

$\leftarrow$

$\rightarrow$

$\leftarrow$

$\rightarrow$

Reveal phase

We will modify this to make players "honest but curious": Bob will try to discover the content of the bit during the commit phase, but would not act on any information he discovers. Quantumly, this means that whenever Alice or Bob does a noisy operation, just do the isometry, skip the partial trace. This means that, if the committed bit is $A$ and $B$ is whatever systems are introduced during the player's attempt to discover or cheat, the global state of $AB$ is always pure.

At commit phase, state is $|\psi_b\rangle_{AB}$ for $b \in \{0,1\}$. Suppose that protocol is perfectly hiding. $\implies \psi_0^B = \psi_1^B$. $|\psi_0\rangle_{AB} = (W \otimes I)\,|\psi_1\rangle_{AB}$. $\implies$ not at all binding. Done.

As a generalizatino, can we have a protocol that is $\epsilon$-hiding, $\delta$-valid? IE Bob can only learn $\epsilon$ information, Alice gets caught with probability $1-\delta$. We would then need need a robust version of purification uniqueness: If $\psi_A \sim \gamma_A$, does there exist $W$ s.t. $\langle\psi|\,(I \otimes W)\,|\gamma\rangle \sim 1$? And that leads us to norms.

**Norms**

A *Norm* is a function $||\cdot|| : x \mapsto ||x||$. Such that:

1. $||cv|| = |c|||v||$ for scalar $c$.
2. $||v + w|| \leq ||v|| + ||w||$
3. $||v|| = 0 \leftrightarrow v = 0$ (separating).

In effect, norms measure the distance between states or matrices. We should think of them as generalizations of the overlaps $||\,\langle\phi|\,|\psi\rangle\,||$ in regular quantum mechanics.

$L_p$ norms are widely used for vectors:

- $||x||_{L_p} = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$
- $L_2 =$ Euclidean
- $L_1 =$ Manhattan
- $L_\infty = max_i |x_i|$

For matrices, we have $S_p =$ Schatten-$p$ norms:

- $||X||_{S_p} = ||svals(X)||_{L_p}$
- $||X||_{S_1} =$ sum of svals = "trace norm"
- $||X||_{S_\infty} =$ biggest singular value

Note that every formulation of quantum mechanics comes with its own natural norm/geometry:

- Pure-state quantum mechanics: $L_2$ unit sphere.
- Probabilities: non-negative vectors in $L_1$ unit sphere
- Density matrices: Positive semi-definite matrices in $S_1$ unit sphere
- Measurement operators: life in $S_\infty$

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 3: Sep 8, 2020

*Lecturer: Aram Harrow* *Scribe: Zane Rossi, Andrey Boris Khesin*

3. Norms, trace distance, fidelity, Uhlmann's theorem

### 3.0.1 Norms

Given a tuple $x \in \mathbb{C}^d$ we define

$$\|x\|_{\ell_p} \equiv \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p} = \|x\|_p,$$

is known as the $\ell_p$ norm. For $\ell_1$ this is useful for probability distributions (where the norm is required to be 1) while $\ell_2$ is useful for pure quantum states (where the norm of a valid state vector is again required to be 1). The $\ell_\infty$ norm finds use when describing classical observables.

We can also define norms over matrices

$$\|M\|_{S_p} \equiv \|\text{sing. val. of } M\|_{\ell_p},$$

is known as the *Schatten p-norm.* Note that the $\|M\|_{S_\infty}$ is simply the largest of the absolute values of the singular values of $M$.

This is relevant for measurements, e.g., $\{M, I - M\}$ is a legal set of measurements iff $0 \leq M \leq I$ in the *PSD ordering* iff $M$ is PSD and its Schatten-$\infty$ norm is less than or equal to 1. Note that $A \geq B$ iff $A - B$ is PSD (possitive-semi-definite).

We can consider the Schatten-2 norm, which is nice operationally,

$$\|M\|_{S_2} \equiv \|\text{vec}(M)\|_{\ell_2} = \sqrt{\text{tr}(M^\dagger M)},$$

We can also consider the Schatten-1 norm,

$$\|M\|_{S_1} \equiv \sum_i |\lambda_i| = \text{tr}(M),$$

where the last equality applies only if the singular values are positive (this means that $M$ is PSD). This is equivalently $\text{tr}\sqrt{(M^\dagger M)}$ if $M$ is a Hermitian operator. All of these results can be seen by taking singular value decompositions of the operator $M$.

The reductions given above are part of a more general scheme of defining inner products over matrices, e.g.,

$$\mathrm{tr} A^\dagger B = \sum_{ij} A^*_{ij} B_{ij} = \langle A, B \rangle.$$

This apparently direct analogy between $\ell_p$ and $S_p$ norms is often useful, and can be a good technique for visualization.

We define the *dual norm* as the result, given a norm $\|\cdot\|_*$, the norm

$$\|x\|_{\mathrm{dual},*} \equiv \max_{\|y\|_* \leq 1} |\langle x, y \rangle|.$$

The $\ell_2$ norm is dual to itself. We can establish

$$|\langle x, y \rangle| \leq \|x\|_{\ell_2} \|y\|_{\ell_2} \leq \|x\|_{\ell_2},$$

while the maximum is achieved by $y = x/\|x\|_{\ell_2}$. This establishes duality of the norm with itself. Note that the direct analogy to this choice does not always hold for other norms.

Note that, under our level of rigour, the dual operation is its own inverse.

We can summarize dualities between $\ell_p$ norms by the following relation,

$$\|\cdot|_{\ell_p,\mathrm{d}} = \|\cdot|_{\ell_q},$$

iff $1/p + 1/q = 1$, which is closely related to the Hölder inequality.

For matrices we use the Hilbert-Schmidt inner product

$$\langle A, B \rangle = \mathrm{tr}(A^\dagger B) = \langle \mathrm{vec}(A), \mathrm{vec}(B) \rangle.$$

We also present the following fact without proof.

$$\|M\|_{S_1} = \max_{U \mathrm{unitary}} |MU|$$

### 3.0.2 Comparing probability distributions

Total variation distance

$$T(p, q) = \frac{1}{2} \|p - q\|_1 = \max_S \left[ p(S) - q(S) \right],$$

where this maximum is taken over subsets of elements,

$$\sum_x |p(x) - q(x)|.$$

Sometimes we want something that looks like the $\ell_2$ norm, though, because of its nice, seemingly geometric properties, but the standard inner product is not great: summing $p(x)q(x)$ over events is generally less than 1.

We can instead consider the fidelity

$$\langle\sqrt{p},\sqrt{q}\rangle = \sum_x \sqrt{p(x)q(x)} = F.$$

Properties of this include $1-F \leq T \leq \sqrt{2(1-F)}$, as well as $F(p_1\otimes p_2, q_1\otimes q_2) = F(p_1,q_1)F(p_2,q_2)$, which is a much nicer behavior than how $T$ acts over tensor products. We also note $F(p,q)\leq 1$ and attains equality iff $p=q$.

These facts help us to prove asymptotic statements like the one below

$$1 - T(p^{\otimes n}, q^{\otimes n}) \sim e^{-cn},$$

where $e^{c_1 n} \leq 1 - T \leq e^{c_2 n}$

### 3.0.3 Comparing quantum states

The naive natural distance here is the $\ell_2$ norm

$$\||\alpha\rangle - |\beta\rangle\|_{\ell_2} = \sqrt{2(1-\mathrm{Re}(\langle\alpha|\beta\rangle))},$$

but this ignores phase and can allow identical states to appear far from each other. Ignoring this we recover the familiar $|\langle\alpha|\beta\rangle|$, which resolves this ambiguity.

For density matrices the trace distance seems natural

$$T(\rho,\sigma) = \frac{1}{2}\|\rho-\sigma\|_{S_1},$$

which is the maximum over 2-outcome measurements $M$ of $\mathrm{tr}(M(\rho-\sigma))$.

We can ask questions now like what happens if we apply a noisy quantum operation?

$$T(N(\rho),N(\sigma)) \quad \text{vs} \quad T(\rho,\sigma)?$$

The easiest way to think about this, says the lecturer, is an isometry followed by a partial trace. Here $N(\rho) = \mathrm{tr}_E(V\rho V^\dagger)$, where $V$ is an isometry iff $V^\dagger V = I$. We know that the isometry cannot change the trace distance, as it implies a corresponding change in the optimal measurement $M \mapsto V^\dagger M V$. The trace over the environment can certainly not increase the chance of distinguishing the two states, and a proof might stem from considering the natural ideal measurement $I\otimes M$ on the entire system with respect to the ideal measurement $M$ on the relevant (i.e., non-traced over) system.

The lecturer makes a comment on another proof of this fact relating operator norms, e.g., that we might alternatively measure $\|N\|_{S_1 \to S_1} \leq 1$.

We can also consider the *mixed state fidelity* or Schatten-1 norm

$$\|\sqrt{\rho}\sqrt{\sigma}\|_{S_1} = \mathrm{tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\sigma}},$$

which is used incredibly rarely. This can make some more natural sense if $\sigma$ is a pure state. Note also that this definition reduces to $|\langle\psi|\phi\rangle|$ in the case of two pure states, and thus might reasonably generalize fidelity to mixed states.

On the problem set we will show $T^2 + F^2 = 1$ for pure states. We might again ask what happens to this measure if we act on it with a quantum channel:

$$F(\rho, \sigma) \leq F(N(\rho), N(\sigma)).$$

**Theorem 3** *Uhlmann's theorem: $F(\rho, \sigma) = \max(|\langle\alpha|\beta\rangle|)$ over $|\alpha\rangle_{AB}$ and $|\beta\rangle_{AB}$ which agree with $\rho$ and $\sigma$ by partial trace over subsystem A (there are many such purifications possible).*

Note that if $|\Gamma\rangle = |i\rangle_A \otimes |i\rangle_B$ (summation implied) is the maximally entangled state, then $\mathrm{tr}_B(\Gamma) = I_A$ the maximally mixed state (unnormalized).

We also introduce the canonical purification $|\phi^\rho\rangle = (\sqrt{\rho} \otimes I_B)\,|\Gamma\rangle$. We can check that this is both normalized and a valid purification. This gives us an equivalent formulation of Uhlmann's theorem.

**Theorem 4** *The alternative form of Uhlmann's theorem is given with proof below:*

$$\begin{aligned}
F(\rho, \sigma) &= \max_U |\langle\phi^\rho| I \otimes U |\phi^\rho\rangle| \\
&= \max_U |\langle\Gamma|(\sqrt{\rho} \otimes I)(I \otimes U)(\sqrt{\sigma} \otimes I)|\Gamma\rangle| \\
&= \max_U |\langle\Gamma|(\sqrt{\rho}\sqrt{\sigma} \otimes U)|\Gamma\rangle| \\
&= \max_U |\mathrm{tr}\sqrt{\rho}\sqrt{\sigma}U^T| \\
&= \|\sqrt{\rho}\sqrt{\sigma}\|_1 \\
&= F
\end{aligned}$$

There are many corollaries to this result, including the Fuch's von-Graf inequalities

$$1 - T \leq F \leq \sqrt{1 - T^2}$$

as well as stronger results against the no-go theorem for quantum bit commitment.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 4: Sep 10, 2020

*Lecturer: Aram Harrow* *Scribe: Annie Wei, Zane Rossi*

## 4.1 Information Theory

Classical information theory

- Shannon entropy, typical sets, and compression
- Mutual information and noisy channel coding
- Relative entropy and hypothesis testing

Quantum information theory

- von Neumann entropy, Schumacher-Jozsa compression
- Mutual information and HSW coding
- Relative entropy and hypothesis testing
- Quantum capacity and LSD theorem

### 4.1.1 Entropy

For random variable $X \sim p$:

$$H(X) = H(p) = -\sum_x p(x) \log p(x)$$

Quantifies uncertainty: for $d$ the alphabet size of $X$,

$$0 \leq H(X) \leq \log d,$$

where lower bound corresponds to deterministic $p = (0, 0, 1, 0, 0)$, upper bounds corresponds to uniform $p = (1/d) \cdot (1, 1, 1, 1, 1)$. Note $0 \log 0 = 0$.

Note: $\ell_\alpha$ norms work also, i.e.

$$\|p\|_{1+\epsilon} = 1 - \epsilon H(p) + O(\epsilon^2)$$

But $\|p\|_0$, $\|p\|_2$, $\|p\|_\infty$ also valid.

In the case of binary entropy, for $\Pi \in [0,1]$,

$$H_2(\Pi) = H\begin{pmatrix} \Pi \\ 1-\Pi \end{pmatrix}$$

### Convexity Properties

Note that $H(p)$ is concave:

$$H(\Pi p + (1-\Pi)q) \geq \Pi H(p) + (1-\Pi)H(q)$$

This inequality is maximized by the uniform distribution. For example, assume that $(0.51, 0.49)$ maximizes entropy. Then $(0.49, 0.51)$ also does. But $H(\text{uniform}) \geq (1/2)H((0.51, 0.49)) + (1/2)H((0.49, 0.51))$.

We can also consider the convexity/concavity properties of fidelity and trace distance. In particular, fidelity is jointly concave:

$$F(\Pi\rho_1 + (1-\Pi)\rho_2, \Pi\sigma_1 + (1-\Pi)\sigma_2) \geq \Pi F(\rho_1, \sigma_1) + (1-\Pi)F(\rho_2, \sigma_2)$$

Trace distance is jointly convex:

$$T(\Pi\rho_1 + (1-\Pi)\rho_2, \Pi\sigma_1 + (1-\Pi)\sigma_2) \leq \Pi T(\rho_1, \sigma_1) + (1-\Pi)T(\rho_2, \sigma_2)$$

To see why this is true, define

$$\begin{aligned} \rho^{AB} &= \Pi \left|1\right\rangle \left\langle 1\right| \otimes \rho_1 + (1-\Pi) \left|2\right\rangle \left\langle 2\right| \otimes \rho_2 \\ \sigma^{AB} &= \Pi \left|1\right\rangle \left\langle 1\right| \otimes \sigma_1 + (1-\Pi) \left|2\right\rangle \left\langle 2\right| \otimes \sigma_2 \end{aligned}$$

Then use the fact that

$$\begin{aligned} F(\rho, \sigma) &= \Pi F(\rho_1, \sigma_1) + (1-\Pi)F(\rho_2, \sigma_2) \\ T(\rho, \sigma) &= \Pi T(\rho_1, \sigma_1) + (1-\Pi)T(\rho_2, \sigma_2) \end{aligned}$$

to get the right hand side of the inequalities. The left hand side comes from

$$\begin{aligned} \rho^B &= \Pi\rho_1 + (1-\Pi)\rho_2 \\ \sigma^B &= \Pi\sigma_1 + (1-\Pi)\sigma_2 \end{aligned}$$

**Joint and Conditional Entropies**

For $X, Y \sim p(x, y)$, define joint entropy

$$H(XY) = H(p) = -\sum_{xy} p(x, y) \log p(x, y)$$

and conditional entropy

$$H(Y|X) = \sum_{x} p(X = x) H(Y|X = x)$$

For a classical distribution $p^{XY} = \Pi_1 |1\rangle \otimes p_1 + \Pi_2 |2\rangle \otimes p_2$,

$$\begin{aligned} H(Y|X = 1) &= H(p_1) \\ H(Y|X = 2) &= H(p_2) \\ \Rightarrow H(Y|X) &= \Pi_1 H(p_1) + \Pi_2 H(p_2) \end{aligned}$$

Note that we can rewrite the conditional entropy as

$$\begin{aligned} H(Y|X) &= -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x) \\ &= -\sum_{xy} p(x) \cdot \frac{p(x, y)}{p(x)} \cdot \log \frac{p(x, y)}{p(x)} \\ &= -\sum_{xy} p(x, y) \log p(x, y) + \sum_{xy} p(x, y) \log p(x) \\ &= H(XY) + \sum_{x} p(x) \log p(x) \\ H(Y|X) &= H(XY) - H(X) \end{aligned}$$

Note also that

$$H(Y|X) \geq 0 \Leftrightarrow H(XY) \geq H(X)$$

although this is not always true quantumly. Also,

$$H(Y|X) \leq H(Y)$$

This statement, that conditioning reduces entropy, is also true quantumly. Note that it's also equivalent to concavity of entropy since

$$\begin{aligned} H(Y|X) &= \Pi_1 H(p_1) + \Pi_2 H(p_2) \\ H(Y) &= H(\Pi_1 p_1 + \Pi_2 p_2) \end{aligned}$$

### 4.1.2 Application: Compression

Say $X \sim p$, and $X^n = (x_1, x_2, ..., x_n) \sim p^{\otimes n}$ are iid samples from $p$. Can I compress $X$?

To do so with 0 error we need $\lceil \log |\text{supp}(p)| \rceil = \log \|p\|_0$ bits. To do so with $\epsilon$ error we need to throw away the smallest elements of $p$ up to weight $\epsilon$.

**Shannon's Noiseless Coding Theorem**

$X^n \sim p^{\otimes n}$, can compress to $n(H(X) + \delta)$ bits with error $\epsilon$ s.t. $\epsilon, \delta \to 0$ as $n \to \infty$.

The converse states that we can't do better. Compressing to $n(H(X) - \delta)$ bits means $\epsilon \to 1$.

Define a **typical set**:

$$T^n_{p,\delta} = \left\{ x^n = (x_1, ..., x_n), \left| -\frac{1}{n} \log p^{\otimes n}(x^n) - H(X) \right| \leq \delta \right\}$$

Define $p^{\otimes n}(x^n) = p(x_1)p(x_2)....p(x_n)$, then

$$\log p^{\otimes n}(x^n) = \sum_{i=1}^{n} \log p(x_i) \to -nH(p)$$

by the law of large numbers. This comes from the fact that

$$E[\log p(x_i)] = \sum_{x_i} p(x_i) \log p(x_i) = -H(p)$$

Thus by the law of large numbers, for all $\delta > 0$,

$$p^{\otimes n}(T^n_{p,\delta}) \to 1$$

as $n \to \infty$. Specifically, for $x^n \in T^n_{p,\delta}$,

$$\exp(-n(H(X) + \delta)) \leq p^{\otimes n}(x^n) \leq \exp(-n(H(X) - \delta))$$

and

$$p^{\otimes n}(T^n_{p,\delta}) \exp(n(H(X) - \delta)) \leq |T^n_{p,\delta}| \leq \exp(n(H(X) + \delta)$$

where the upper bound is used in the coding theorem, and the lower bound is used in the converse. Thus the number of bits needed is

$$\log |T^n_{p,\delta}| \leq n(H(X) + \delta)$$

Next time we'll look at Shannon's noisy coding theorem.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 5: Sep 15, 2020

*Lecturer: Aram Harrow* *Scribe: Changnan Peng, Annie Wei*

# 5.1 Information Theory: Classical and Quantum

## 5.1.1 Noiseless Coding Theorem

### Review

Today we will continue talking about information theory.

Recall that last time we defined the **Shannon entropy** as a measure of uncertainty for the probability distribution $p(x)$:

$$H(p) = -\sum_x p(x) \log p(x). \tag{5.1}$$

Then we discussed an application of Shannon entropies to the problem of compression. Recall that we started by defining a **typical set** (also known as the ”asymptotic equipartition property”), which is a set of strings with probability $\delta$-close to a probability distribution $p$:

$$T^n_{p,\delta} = \left\{ x^n = (x_1, ..., x_n) \text{ s.t. } \left| -\frac{1}{n} \log p^{\otimes n}(x^n) - H(x) \right| \leq \delta \right\}. \tag{5.2}$$

The probability of being non-typical, characterized by $\epsilon = 1 - p^{\otimes n}(T^n_{p,\delta})$, goes to zero, as $n \to \infty$. In fact, $\epsilon \leq n^{O(1)} 2^{-n\delta}$.

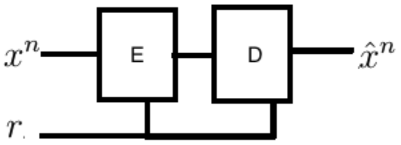
### Converse

Now, continuing from last time, let's talk about the **converse** to the Shannon coding theorem: Let's say that we compress to $k$ bits, and we assume that there exists a set $S$ that is decoded correctly. Note that by definition $|S| \leq 2^k$. Then the probability of correctly decoding is

$$p^n(S) \leq p^n(T^n_{p,\delta} \cap S) + p^n(\overline{T^n_{p,\delta}}) \leq 2^k 2^{-nH(p)+n\delta} + \epsilon \tag{5.3}$$

Note that the right hand side is a small number if $k \leq n(H(p) - \delta)$. In this argument we assumed deterministic coding and encoding, but what if Alice and Bob share randomness? We'll claim that even with the shared randomness, the bound still holds.

Note that with shared randomness, we have a diagram that looks something like the following:



Note that we can condition on the shared randomness $r$, reducing it to the deterministic case, and then sum over $r$,

$$P(x^n = \hat{(x)}^n | r) \leq \epsilon', \tag{5.4}$$

so shared randomness should not change our results.

### 5.1.2 Quantum Entropy and Compression

#### Quantum Entropies

In the quantum case, we can define the von Neumann entropy,

$$S(\rho) = H(\text{eig}(\rho)) = -tr[\rho \log \rho]. \tag{5.5}$$

It is zero if and only if $\rho$ is a pure state:

$$S(\rho) = 0 \Leftrightarrow eig(\rho) = (1, 0, ..., 0) \Leftrightarrow \rho = |\psi><\psi|. \tag{5.6}$$

It attains its maximum when $\rho$ is maximally mixed. Letting $d = \dim(\rho)$,

$$S(\rho) \leq \log d \tag{5.7}$$

$$S(\rho) = \log d \Leftrightarrow \rho = I/d. \tag{5.8}$$

We can generalize all of our classical entropies to the quantum case described by density matrix $\rho_{XY}$:

$$S(X) = S(\rho_X) \tag{5.9}$$

$$S(X|Y) = S(XY) - S(Y) \tag{5.10}$$

$$I(X:Y) = S(X) + S(Y) - S(XY) \tag{5.11}$$

$$D(\rho||\sigma) = \mathrm{tr}\rho[\log\rho - \log\sigma] \tag{5.12}$$

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{5.13}$$

Note that again $S(X|Y) \leq S(X)$, which allows us to derive concavity of entropy. A good question to ask is how we should actually interpret the quantity $S(X|Y)$. For example, for the state

$$|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2},$$

how would we actually condition on the state of the second system? This isn't so clear! Note, also that the conditional entropy in the quantum case can be negative. An example is again the Bell state, where $S(XY)_\psi = 0$, $S(Y)_\psi = 1$, $S(X|Y) = -1$.

Now let's extend the notion of typical sets. Say we have the state

$$\rho = \sum_x \lambda_x |v_x\rangle\langle v_x|,$$

where the $|v_x\rangle$'s are orthonormal. Then we can define a typical projector

$$\Pi^n_{p,\delta} = \sum_{x^n \in T_{\lambda,\delta}} |v_{x^n}\rangle\langle v_{x^n}|.$$

Here

$$|v_{x^n}\rangle = |v_{x_1}\rangle \otimes ... \otimes |v_{x_n}\rangle.$$

This results in

$$\rho^{\otimes n} = \sum_{x^n} \lambda_{x_1}...\lambda_{x_n} |v_{x^n}\rangle\langle v_{x^n}|$$

Note that this projector projects to the typical subspace. The projection measurement $\{\Pi^n_{p,\delta}, I - \Pi^n_{p,\delta}\}$ has resulting probability

$$\mathrm{tr}[\rho^{\otimes n}\Pi^n_{p,\delta}] = \sum_{x^n} \lambda_{x_1}...\lambda_{x_n} 1_{x^n \in T^n_{\lambda,\delta}} = \lambda^n(T^n_{\lambda,\delta}) \tag{5.14}$$

Note that this approaches 1 as $n \to \infty$.

Note that we would like to discuss compressing unknown $\rho$ for qubits with known $S(\rho)$. If $\rho$ is known, then we can use a classical compression scheme working in the eigenbasis $|v_1\rangle$,...,$|v_d\rangle$.

### Efficient Classical Compression

Now let's look at an example of an **efficient classical compression scheme**, specifically Huffman encoding.

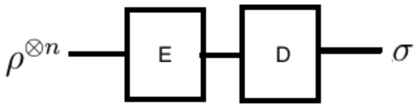
For an example case, suppose we have 4 symbols, with probabilities given in the second column below. Then we can assign a code using the third column below:

| | | |
|---|---|---|
| $A$ | 1/2 | 0 |
| $B$ | 1/4 | 10 |
| $C$ | 1/8 | 110 |
| $D$ | 1/8 | 111 |

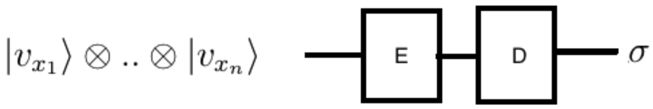
Note that this encoding encodes $x$ with $\lceil \log 1/p(x) \rceil$ bits, and that this is always possible. Note also that this encoding is prefix-free. If the probabilities are not powers of 1/2, we can use block coding, i.e. by expanding our code to include multiple bits in a block. In our example, we would take the codewords to be blocks $AA, AB, BB$, etc. Then this allows us to assign probabilities to each block that are closer to powers of 1/2.

### Quantum compression

We start off by asking the question, what does it mean to compress $\rho$? Here are some possible answers:

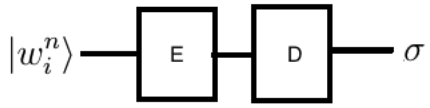
1. 

   with $F(\rho^{\otimes n}, \sigma) \approx 1$.

2. 

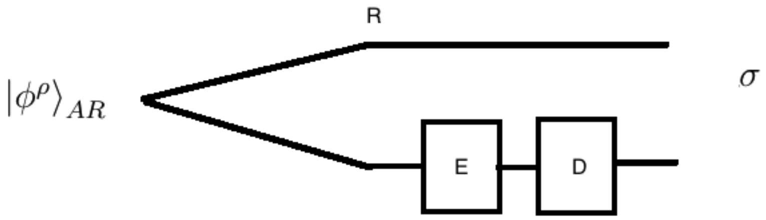
with $E_{x^n \sim \lambda}[F(|v_{x^n}\rangle \langle v_{x^n}|, \sigma)] \approx 1$.

3. Letting $\rho = \sum_i p_i |w_i\rangle \langle w_i|$, where the $|w_i\rangle$ are not necessarily orthonormal,



with $E_{i^n \sim \rho}[F(|w_{i^n}\rangle, \sigma) \approx 1]$.

4. Letting $|\phi^\rho\rangle_{AR}$ be a purification such that $\phi^\rho_A = \rho$,



with $F(|\phi^\rho\rangle^{\otimes n}, \sigma) \approx 1$.

Note that the first scheme doesn't work because it allows for the possibility where you input the maximally mixed state and produce the output by just throwing away the input state and always outputting the maximally mixed state. The classical equivalent would be a source that always emits the uniform distribution, and an encoding scheme that throws away the acutal message and just returns the uniform distribution.

The second to fourth options are roughly the same, and give us **Schumacher Jozsa compression**. Formally, the way this works is the following: Say we have state $\rho^{\otimes n}$. Apply $\{\Pi^n_{p,\delta}, I - \Pi^n_{p,\delta}\}$, where the first result is successful and the second is a failure. If this is successful, the state is contained in a subspace of dimension $\mathrm{tr}\Pi^n_{p,\delta} \leq \exp(n(S(\rho) + \delta) = \exp(nH(\lambda) + \delta)$. Thus it fits into $n(S(\rho) + \delta)$ qubits.

An application of this is to algorithmic cooling, where our states (representing, for example, nuclear spins in large magnetic fields) are of the form $\rho^{\otimes n}$ with

$$\rho = \begin{pmatrix} \frac{1+\epsilon}{2} & 0 \\ 0 & \frac{1-\epsilon}{2} \end{pmatrix}.$$

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 6: Sep 17, 2020

*Lecturer: Aram Harrow* *Scribe: Andrew Tan, Changnan Peng*

## 6.1 Relative Entropy

In the previous lectures, we introduced information entropy. An alternative interpretation of entropy is as "average surprise". In our daily experience, a more likely event contains less information and brings us less surprise. For example, if the weather forecast said there would be 90% probability of raining and it rains, we would not be very surprised. If it said 10% and rains, we would be more surprised. It is similar for the events in Huffman coding. We would be more surprised when an event with probability $2^{-10}$ appears than when one with $2^{-1}$ does. Huffman coding offers us a quantification of surprise. Given $n$ bits, we can identify one of $2^n$ events each with probability $2^{-n}$. This suggests that an event with probability $p(x)$ need $\log \frac{1}{p(x)}$ bits.

We can define

$$\text{surprise}(x) \equiv \log \frac{1}{p(x)} \tag{6.1}$$

and therefore the "average surprise"

$$\mathbb{E}[\text{surprise}(x)] = \sum_x p(x) \log \frac{1}{p(x)} = H(p) \tag{6.2}$$

Also in Huffman coding, $x$ uses $\lceil \text{surprise}(x) \rceil$ bits. That's how entropy as "average surprise" measures information.

In the previous example of Huffman coding, we have assumed that we know the true distribution of the events and encode them accordingly. What if we use the wrong distribution (i.e. $x \sim p$, but we encode according to $q$)? For example, we compress a piece of text by encoding the letters according to their appearance probability in English, but actually the text is written in French. In such cases, we cannot have optimal compression. The compressed message is longer than the one encoded with the correct distribution. The message length $\sum_x p(x) \log \frac{1}{q(x)} \geq \sum_x p(x) \log \frac{1}{p(x)}$.

The excess, denoted $D(p||q)$ is known as the *relative entropy* or *Kullback-Leibler divergence*

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{6.3}$$

The relative entropy is always non-negative. One can see this roughly by noting that Shannon's coding theorem implies that we cannot compress a source beyond its entropy, and therefore the excess must be $\geq 0$. However, this conclusion is not obvious from viewing the formula. Unlike in the definition of entropy where every term is non-negative, here the terms have mixed signs, being non-negative on the support where $p(x) \geq q(x)$, and negative otherwise. The non-negativity of relative entropy comes from the positive terms outweighing the negative ones.

Showing this more rigorously, we make use of the fact that

$$1 + z \leq e^z,$$

which can be shown from the convexity of $f(z) = e^z - (z+1)$, and $f(0) = f'(0) = 0$.

Replacing $z$ by $\log y$, we get the equivalent forms

$$\begin{aligned} \log y &\leq y - 1 \\ \log \frac{1}{y} &\geq 1 - y. \end{aligned}$$

Applying this inequality,

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\geq \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) \\ &= \sum_x p(x) - q(x) = 0, \end{aligned}$$

in the last step we used $\sum_x p(x) = \sum_x q(x) = 1$.

From the definition of relative entropy, we can see that it is zero when $p = q$. Are there any other cases? Tracing through the derivation of non-negativity, the inequality is tight only at one point, $z = 0$, equivalently $y = 1$ or $p(x) = q(x)$. A little tricky point is that at the terms with $p(x) = 0$, the inequality may also be tight, as those terms are zero. However, $\sum_x p(x) = \sum_x q(x) = 1$ forces $q(x)$ to be 0 when $p(x) = 0$, given $p(x) = q(x)$ when $p(x) \neq 0$.

Therefore, $D(p||q) = 0$ if and only if $p = q$.

Another note is that although the relative entropy describes the difference between two distributions, it is not a true distance in a metric sense – it is neither symmetric, $D(p||q) \neq D(q||p)$, nor satisfying the triangular inequality.

### 6.1.1 Corollary: Subadditivity of entropy

Using the non-negativity of relative entropy, we can prove the subadditivity of information entropy. Consider a joint distribution $p_{XY}$, and the direct product of its marginals, $p_X \otimes p_Y$. We calculate the relative entropy between them, and we can group the terms in different ways.

$$\begin{aligned}
D(p_{XY}||p_X \otimes p_Y) &= \sum_{x,y} p(x,y)\left(\log p(x,y) - \log p_X(x) - \log p_Y(y)\right)\\
&= -H(XY) + H(X) + H(Y)\\
&= H(X) - H(X|Y)\\
&= H(Y) - H(Y|X)\\
&\equiv I(X:Y) \geq 0.
\end{aligned}$$

The second line tells us the subadditivity of entropy, i.e. $H(X)+H(Y) \geq H(XY)$. The third and fourth lines tell us conditioning on other systems will decrease the entropy, i.e. $H(X) \geq H(X|Y)$ and $H(Y) \geq H(Y|X)$.

In the last line we introduce a new quantity $I(X:Y)$, known as the *mutual information*. It describes the correlation of $X$ and $Y$ in a joint distribution $p_{XY}$. The mutual information $I(X:Y) = 0$ if and only if $X$ and $Y$ are independent, i.e. $p_{XY} = p_X \otimes p_Y$.

### 6.1.2 Corollary: Uniform distribution has largest entropy

We can also use the non-negativity of relative entropy to show that the uniform distribution has the largest entropy. Consider the special case of a distribution $p$ with the uniform distribution $u = \left(\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}\right)$ on $d$ outcomes,

$$\begin{aligned}
D(p||u) &= \sum_x p(x)\left(\log p(x) - \log\frac{1}{d}\right)\\
&= \log d - H(p) \geq 0.
\end{aligned}$$

It tells us $H(p) \leq \log d$, and this maximum is reached if and only if $p = u$.

## 6.2 Hypothesis testing

The key application to understand information entropy is the message compression. We will see in this section that the key application for the relative entropy is the hypothesis testing.

In a hypothesis testing, we are given several hypotheses of distributions and a sample of data. We would like to find out which distribution the sample comes from. When two hypotheses are given, it is called *binary hypothesis testing*. When there are more than two hypotheses, it is *multiple hypothesis testing*. Here we only talk about binary hypothesis testing.

Suppose we get $x$ sampled from $p$ or $q$ and want to guess which distribution $x$ comes from. There are two kind of errors we can make – $x$ sampled from $p$ but we guess $q$ (type 1), or $x$ sampled from $q$ but we guess $p$ (type 2). We define the probability of these two types of errors as

$$\begin{aligned} \alpha &= \Pr[\text{guess } q|x \sim p] \qquad &\text{(type 1)} \\ \beta &= \Pr[\text{guess } p|x \sim q] \qquad &\text{(type 2)} \,. \end{aligned}$$

We want to do the hypothesis testing that can minimize these errors. There are several ways to formulate the problem

1. Symmetric hypothesis testing: minimize $\alpha + \beta$. Answer is $\|p - q\|$.
2. Bayesian hypothesis testing: minimize $\pi\alpha + (1 - \pi)\beta$. Answer on problem set 1.
3. Asymmetric hypothesis testing: minimize $\beta$ such that $\alpha \leq \epsilon$. Minimum is $\beta_\epsilon$.

As usual in the information theory, we consider the asymptotic case of $n$-copies with $n \to \infty$. Intuitively, with more samples in hand, we can distinguish the distributions better. When we cap $\alpha$ by a fixed value, $\beta$ should decrease exponentially with $n$. The question left is the coefficient in front of $n$ in the exponent, and the answer is the relative entropy.

Define $\beta_\epsilon^n$ to be the minimum of type-2 error for the binary hypothesis testing between $p^{\otimes n}$ and $q^{\otimes n}$. We expect $\lim_{n\to\infty} \beta_\epsilon^n \sim \exp\left(-nD(p||q)\right)$. Formally, we have the following theorem.

**Theorem 5 (Chernoff-Stein's Lemma)**

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_\epsilon^n = D(p||q), \quad \forall \epsilon \in (0, 1). \tag{6.1}$$

We will not give a proof here. Instead, let's see some examples.

1. $p = q \iff D(p||q) = 0$. It is obvious that we cannot distinguish two distributions when they are identical. On the other direction, when two distributions are different, no matter how they are alike, we can do the hypothesis testing with exponentially small error given large enough number of samples.

2. $q$ is the uniform distribution $u$. $D(p||u) = \log d - H(p)$. We can do the following hypothesis testing. If the samples are in the typical set of $p$, i.e. $x^n \in T^n_{p,\delta}$, we guess $p$, otherwise we guess $q$. The appearance of the type-1 error is when $p$ generate samples outside the typical set. From the property of the typical set, we know the probability of type-1 error is $\alpha = 1 - p^{\otimes n}(T^n_{p,\delta}) \to 0$, for all $\delta \geq 0$. The appearance of the type-2 error is when the samples generated from the uniform distribution happen to be in the typical set. The probability is $\beta = \frac{|T^n_{p,\delta}|}{d^n} \leq \exp\left(n\left(H(p)+\delta\right) - n\log d\right) \leq \exp\left(-n\left(D(p||u) - \delta\right)\right)$. The minimal error must be small than this, i.e. $\beta^n_\epsilon \leq \beta \leq \exp\left(-n\left(D(p||u) - \delta\right)\right)$. We can smoothly reach the bound stated in the theorem by making $\delta$ slowly goes to zero.

3. $D(p||q) = \infty$. From the definition of the relative entropy, this occurs when there is an element $x$ such that $p(x) \neq 0$ and $q(x) = 0$, i.e. when $\text{supp}(p) - \text{supp}(q) \neq \varnothing$. We can do the following hypothesis testing. If an element in $\text{supp}(p) - \text{supp}(q)$ is seen, we guess $p$, otherwise we guess $q$. The type-1 error appears when those elements happen not to be seen, which occurs with probability that decreases exponentially. Note that we can always guess $p$ with certainty. Therefore, the probability of the type-2 error $\beta = 0$.

## 6.3 Quantum relative entropy

Now let's consider the quantum case. Unlike in the classical case where we can divide two probability distributions, the quantum analog of the relative entropy is defined as follows:

$$D(\rho||\sigma) \equiv \text{tr}\left[\rho\log\rho - \rho\log\sigma\right] = \text{tr}\left[\rho\left(\log\rho - \log\sigma\right)\right] \tag{6.1}$$

There is $D(\rho||\sigma) \geq 0$. A consequence of this is that the quantum mutual information $I(X:Y) \equiv D(\rho_{XY}||\rho_X \otimes \rho_Y) \geq 0$ is still non-negative by the same arguments as in the classical case.

We have a similar theorem for the binary hypothesis testing in the quantum case.

**Theorem 6 (Quantum Stein's lemma)** *Given $\rho^{\otimes n}$, $\sigma^{\otimes n}$. For any possible two-outcome measurement $\{M, 1-M\}$, define the minimal type-2 error given a capped* *type-1 error*

$$\beta_\epsilon^n = \min\left\{\mathrm{tr}[M\sigma^{\otimes n}] \quad | \quad \forall M\,\mathrm{s.\,t.}\,\mathrm{tr}[M\rho^{\otimes n}] \geq 1-\epsilon\right\}.$$

*The following limit holds*

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_\epsilon^n = D(\rho||\sigma).$$

We will not give a proof here. Instead, we consider the special case $D(\rho||\sigma) = \infty$. This time there is no clear meaning of probability on an element as in the classical case. Instead, the support is defined as the span of all eigenvectors. In the quantum case, $D(\rho||\sigma) = \infty \iff \mathrm{supp}\,\rho \not\subseteq \mathrm{supp}\,\sigma$.

When $\rho$ and $\sigma$ are pure states, e.g. $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$. The support of $\rho$ is $\{|\psi\rangle\}$ and the support of $\sigma$ is $\{|\phi\rangle\}$. This implies that $D(\rho||\sigma) = 0$ for identical pure states, or $D(\rho||\sigma) = \infty$ otherwise. Therefore, the relative entropy is not a well description for the difference between two pure states.

The optimal measurement in this case, is to choose $M = |\psi^\perp\rangle\langle\psi^\perp|$ and $1-M$. Note that the optimal measurement is not parallel to the state, but instead perpendicular. With this measurement we can rule out one of the hypothesis definitely. The proof is similar as in the classical case.

### 6.3.1 Quantum versus classical entropies, Conditional mutual information

Here are some properties of the quantum entropy

1. $0 \leq S(X) \leq \log d$ with equality on the lower bound only for pure states and equality for the upper bound only for the maximally mixed state $I/d$.
2. $0 \not\leq S(X|Y) \leq S(X)$. the non-negativity of the conditional entropy only holds in the classical case.
3. $D(\rho||\sigma) \geq 0$
4. $I(X:Y) \geq 0$

There is another quantity we have not yet introduced in this family. The *conditional mutual information* is the amount of mutual information conditioned on another random variable. It combines the idea of conditional entropy and mutual information.

Classically,

$$I(X:Y|Z) = \sum_z p_z(z) I(X:Y)_{p(\cdot,\cdot|z)} \geq 0 \tag{6.2}$$

and $I(X:Y|Z) \geq 0$ follows directly from subadditivity. The following equivalent definitions hold in both the classical and quantum cases

$$\begin{aligned} I(X:Y|Z) &= H(X|Z) + H(Y|Z) - H(XY|Z) \\ &= H(XZ) - H(Z) + H(YZ) - H(Z) - H(XYZ) + H(Z) \\ &= H(XZ) + H(YZ) - H(XYZ) - H(Z) \\ &= I(X:YZ) - I(X:Z). \end{aligned}$$

In the quantum case, it is still true that $I(X:Y|Z) \geq 0$ but does not follow obviously from subadditivity. This property is known as the "*strong subadditivity* (SSA) of quantum entropy". The proof is far more complicated than in the classical case.

However, the relation between $I(X:Y|Z)$ and $I(X:Y)$ is not definite. It can be "$\geq$", "$=$", or "$\leq$". For example, if $Z$ describe the noise that is added on both $X$ and $Y$, conditioning on the noise can increase the mutual information between the signals. On the other hand, it is also possible that $Z$ exactly determines $X$ and $Y$. In this case, the conditional mutual information equals zero.

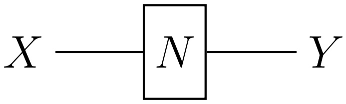
**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 7: Sep 22, 2020

*Lecturer: Aram Harrow* *Scribe: Zeyang Li, Andrew Tan*

# 7.1 Noisy Channel Coding

## 7.1.1 Classical noisy channels



We would like to study communication in a realistic setting where the medium over which messages are transmitted can (partially) corrupt the signal. We model a classical noisy channel $N$ as a mapping between random variables $X$ to $Y$, $N(y|x)$,

$$p_Y(y) = p_X(x)N(y|x)$$

This leads to a natural question: what is the largest amount of information that can be communicated per use of a given channel $N$? This quantity, known as the *channel capacity*, is defined as the highest rate of reliable communication that can be achieved, measured in bits per channel use, over all possible coding strategy; reliability in this case refers to the probability of error $\epsilon \to 0$ in the asymptotic data limit $n \to \infty$.

Mathmatically,

$$C(N) \equiv \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log M^* \tag{7.1}$$

where

$$M^* = \max\left\{M : \exists E : [M] \to X^n, \exists D : Y^n \to [M], \text{s.t.} \forall m, \Pr[m = D(N^{\otimes n}(E(m)))] \geq 1 - \epsilon\right\} \tag{7.2}$$

where the notation $[M] \equiv \{1, 2, \ldots, M\}$.

Shannon's noisy coding theorem provides the answer in simple expression:

$$C(N) = \max_{p_x} I(X : Y)_p \tag{7.3}$$

where $p(x, y) = p_X(x)N(y|x)$

Here the $p$ is the joint input-output distribution. We can understand the mutual information in number of equivalent ways:

$$I(X;Y) = H(X) - H(X|Y)$$

The first interpretation of the mutual information is as the amount of information in the random variable $X$ less the amount of uncertainty in $X$ that still remains after observing the, potentially (partially) corrupted $Y$

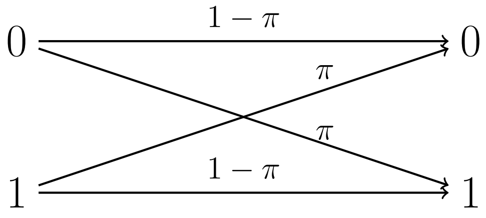
$$I(X;Y) = H(Y) - H(Y|X)$$

Another interpretation is as the amount of information carried in the observed $Y$ less the randomness in $Y$ that carries no information about $X$ (i.e. the noise injected by the channel).

$$I(X;Y) = D(p_{XY}||p_X \otimes p_Y)$$

Yet another interpretation is as the relative entropy between the correlated joint distribution $p_{XY}$ and the independent product of the marginals $p_X \otimes p_Y$

**Example: Binary Symmetric Channel (BSC)**



A commonly studied noisy channel model is the *binary symmetric channel* (BSC). The BSC has binary inputs and outputs with a probability $\pi$ of the sent bit being flipped. That is the output

$$Y = X \oplus e,\ \ e = \begin{cases} 0, & \text{w/ prob. } 1-\pi \\ 1, & \text{w/ prob. } \pi \end{cases}$$

The Shannon limit defined in Equation 7.3, for the BSC is $C(N_{BSC}) = 1 - H_2(\pi)$; where $H_2(\pi)$ is the known as the binary entropy function

$$H_2(\pi) \equiv -\pi \log \pi - (1-\pi) \log(1-\pi)$$

we can see this by noting that the entropy is a concave function of $p_X$ and is symmetric about the mid-point $\pi = \frac{1}{2}$ and is therefore maximized for $p_X = \left(\frac{1}{2}, \frac{1}{2}\right)$ (also note that

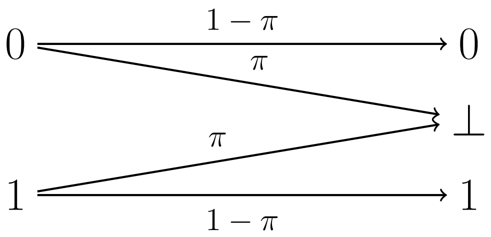
$p_Y = \left(\frac{1}{2}, \frac{1}{2}\right)$). From this, we can obtain the joint distribution

$$p = \frac{1}{2}\begin{pmatrix} 1-\pi & \pi \\ \pi & 1-\pi \end{pmatrix}$$

where we write the joint distribution $p$ in the form of a $2 \times 2$ matrix.

Intuitively, this appears to be the best possible rate since, by correctly decoding the output $Y$, one obtains all of the information that has been input: one bit of information corresponding to the sum of the entropy of $X$ as well as the entropy of the noise $e$.

### Example: Erasure Channel



Another commonly studied channel is the erasure channel where a bit is lost with probability $\pi$. By the same argument as above, the maximum of Equation 7.3 is again attained for $p_X = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Here we have $H(Y|X) = H_2(\pi)$ and $H(Y) = 1-\pi+H_2(\pi)$ giving a channel capacity $C(N_{ERASURE}) = 1 - \pi$.

Once again, this appears to be the best possible rate. One can imagine a protocol where Alice communicates to Bob using the channel $N_{ERASURE}$ and Bob has a noiseless channel to Alice that can be used to confirm the reception of a bit or the loss of a bit. In the case that Bob loses the bit, he communicates this noiselessly to Alice asking for her to send it again. This occurs with rate $1 - \pi$. This appears to be the best-case scenario as it requires access to an unphysical noiseless channel; and it is somewhat surprising that the channel capacity saturates the upper-bound given by this idealized scenario.

### Example: Gaussian Noise Channel

A common model for analog communication is that of the Gaussian noise channel.

For concreteness, consider $x^n \in \mathbb{R}^n$ and $e \sim \mathcal{N}(0, \sigma^2)$. Finding the channel capacity in this case is close to the problem of sphere packing – especially as the Gaussian balls become less 'fuzzy' in high dimensions. Note here that Bob learns $y^n$ and also gets $x^n$ if decoding works, and then therefore reconstruct the noise $e^n$.

**Is the channel capacity achievable?**

From the examples above, it seems clear that the channel capacities in these cases are as high as one could expect; however, it is not immediately clear that we can find an encoding scheme that achieves the channel capacity.

Consider using the repetition code over the BSC: encode $0 \mapsto 0^k, 1 \mapsto 1^k$, and the error rate $\Pr[\text{error}] \sim e^{-\mathcal{O}(k)}$ – this can be shown more rigorously using Chernoff bounds.

Consider the case where Alice would like to send a message of length $l$, with each bit encoded by a $k$-fold repetition. The total length is $n = kl$. The error rate per encoded bit goes as $e^{-k}$. To reliably transmit the entire message, we require the error rate per block to be less than $1/l$ corresponding to a choice of $k \sim \log l$. This gives a rate of $R \sim \frac{1}{\log n}$ showing that as $n \to \infty$, the rate of this encoding scheme goes to zero. We will need to be more clever if we are to achieve the channel capacity.

### 7.1.2 Proof of Shannon's Noisy-Channel Coding Theorem

Now we will prove that the channel capacity defined in Equation 7.3 is achievable by providing constructing a suitable encoding scheme.

First we define the *jointly typical set* $J^n_{p,\delta}$ as follows:

$$J^n_{p,\delta} \equiv \left\{(x^n, y^n) : (x_1y_1, \ldots, x_ny_n) \in T^n_{p_{XY},\delta}, (x_1, \ldots, x_n) \in T^n_{p_X,\delta}, (y_1, \ldots, y_n) \in T^n_{p_Y,\delta}, \right\} \tag{7.4}$$

that is the $J^n_{p,\delta}$ is the set of length $n$ pairs of $(x^n, y^n)$ such that $x^n$ is typical with respect to $p_X$, $y^n$ is typical with respect to $p_Y$ and $(x^n, y^n)$ is typical with respect to $p_{XY}$ simultaneously.

From the results on the typical set, we have the following properties of strings in the jointly typical set

$$\begin{aligned} p^n_{XY}(x^n, y^n) &\approx \exp\left(-nH(XY)\right) \\ p^n_X(x^n) &\approx \exp\left(-nH(X)\right) \\ p^n_Y(y^n) &\approx \exp\left(-nH(Y)\right) \end{aligned}$$

and $p^n_{XY}(J^n_{p,\delta}) \to 1$ as $n \to \infty$.

In encoding process, we have a random codebook, $C = \{E(1), \cdots, E(M)\}$, where $M = |C| = 2^{nR}$, and $R$ is the rate. Each $E(m)$ is drawn independently from $p_x^{\otimes n}$. To decode, we perform *joint typicality decoding*: given output $y^n = N(x^n)$, $D(y^n) = \hat{m}$ s.t. $(E(\hat{m}), y^n) \in J^n_{p,\delta}$. This can fail if

1. $\hat{m}$ does not exist, or
2. $\exists \hat{m} \neq m$ satisfying $(E(m'), y^n) \in J$.

We now show that both of these are unlikely for $R \leq C(N)$. Consider uniform distributed $m \in [M]$:

$$\Pr[\underbrace{(E(\hat{m}), y^n)}_{\sim p_{xy}^{\otimes n}} \in J^n_{p,\delta}] = p_{xy}^{\otimes n}(J^n_{p,\delta}) \to 1 \text{ as } n \to \infty$$

demonstrating that the first failure mode is unlikely.

Since $E(m)$ and $E(m')$ are independent, $E(m')$ and $y^n$ are independently distributed, therefore for a fixed $m'$,

$$\Pr[m' \neq m \cap (\underbrace{E(m')}_{\sim p_x^n}, \underbrace{y^n}_{\sim p_y^n}) \in J^n_{p,\delta}] = (p_x^n \otimes p_y^n)(J^n_{p,\delta}).$$

Since $p^n(x^n, y^n) \geq \exp(-nH(XY) - n\delta)$, we have $|J^n_{p,\delta}| \leq \exp(nH(XY) + n\delta)$, and therefore the r.h.s is

$$\begin{aligned} (p_x^n \otimes p_y^n)(J^n_{p,\delta}) &\leq |J^n_{p,\delta}| \max p_x^n \max p_y^n \\ &= \exp(-nH(X) + n\delta) \exp(-nH(Y) + n\delta) \exp(nH(XY) + n\delta) \\ &= \exp(-nI(X:Y) + 3n\delta) \end{aligned}$$

for a fixed $m'$.

For all $m'$,

$$\Pr[\exists_{m' \neq m} \text{s.t.} (E(m'), y^n) \in J^n_{p,\delta}] \leq M(p_x^n \otimes p_y^n)(J^n_{p,\delta}) \leq \exp(nR - nI(X;Y) + 3n\delta)$$

which → 0 when $R < I(X:Y) - 3\delta$.

Note that this provides an another interpretation of the mutual information $I(X;Y)$.

Although we have proven the achievability of the Shannon limit, the use of random codebook and joint typicality decoding is quite messy. Next class we're going to get rid of the random codebook and random message.

## 7.2 Quantum analogues

### 7.2.1 CQ channel capacity

Consider classical input, quantum ouput or $CQ$ channel $N$. This can be thought of a channel that takes as input a number $x \in [M]$ and outputs a quantum state $\rho_x$; it can also be thought of as a channel that takes a quantum state $\sigma$ but immediately decoheres it:

$$N(\sigma) = \sum_x \langle x|\sigma|x\rangle \, \rho_x$$

What is the classical capacity of this? The answer to this is given by the Holevo-Schumacher-Westmoreland (HSW) theorem:

$$C(N) = \max_p I(X;Q)_\omega \tag{7.1}$$

where $\omega^{XQ} = \sum_x p(x)|x\rangle\langle x|^X \otimes \rho_X^Q$; and the CQ joint entropy is

$$\begin{aligned} S(XQ) &= -\text{tr}\left[\omega^{XQ} \sum_x |x\rangle\langle x| \otimes (\log p(x)I + \log \rho_x)\right] \\ &= H(p) + \sum_x p_x S(\rho_x) = H(X) + H(Q|X) \end{aligned}$$

and the CQ mutual information is

$$\begin{aligned} I(X;Q) &= H(Q) - H(Q|X) \\ &= S\left(\sum_x p(x)\rho_x\right) - \sum_x p(x)S(\rho_x) = \chi \end{aligned}$$

**Example: simple application of the HSW theorem**

Consider qubit states $\rho_i = |v_i\rangle\langle v_i|$ for $i \in \{1,2,3\}$. Assume that they related by $\frac{2\pi}{3}$ rotation so that $\frac{1}{3}(\rho_1+\rho_2+\rho_3) = I_2/2$. In this case, the $\rho_i$ are pure states and therefore $S(\rho_i) = 0$ giving $S = 1$ and $\chi = I(X;Q) = 1$. This means that we can reliably transmit 1 bit of information per use of the channel. This would be clear if the output states were orthogonal such as $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$, but is not as obvious in this case where the output states are not orthogonal.

### 7.2.2 Quantum joint typicality

The quantum analogue for typical sets are projectors into jointly typical subspaces: $T_X \mapsto \Pi_X$, $T_Y \mapsto \Pi_Y$, and $T_{XY} \mapsto \Pi_{XY}$. The problem, however, is that these projectors do not commute in general and therefore we cannot directly define joint typicality.

To get around this, we can first purify the state $\rho_{XY} \mapsto |\psi\rangle_{XYZ}$ and consider $\Pi_Z$ which should be fine as the projectors have the same spectrum. Another way is to consider quantities of the form $\Pi_{XY}\Pi_X\rho^{\otimes n}\Pi_X\Pi_{XY}$ although one needs to be careful of the ordering of the operators. This will be expanded on in the next lectures as we prove the HSW theorem.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 8: Sep 24, 2020

*Lecturer: Aram Harrow* *Scribe: Thao Dinh, Zeyang Li*

## 8.1 Shannon's Noisy Coding Theorem (cont)

Last class's proof we have two key features of Shannon's noisy coding theorem are random encoding and jointly typical decoding. The probability of error averaged over all the messages $m$, codebook $C$, and actions of the channel $N^n$ is small.

$$\Pr_{m,C,N^n}[\text{error}] \leq \epsilon$$

The average is always greater than minimum, and therefore the LHS, i.e., the expectation value over $C$ of the probability of error given the choice of $C$ is greater than the minimum over $C$ of the probability of error.

$$\mathop{\mathbb{E}}_{C}\left[\Pr_{m,N^n}[\text{error}|C]\right] \geq \min_{C} \Pr_{m,N^n}[\text{error}]$$

Fix the codebook $C$ to be the one with minimum probability of error, but here we want it works for all message rather than some particular ones. In this case we can use the Markov's inequality, i.e. given a non negative random variable $X$, the probability to have $X \geq a$ is:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}(X)}{a}$$

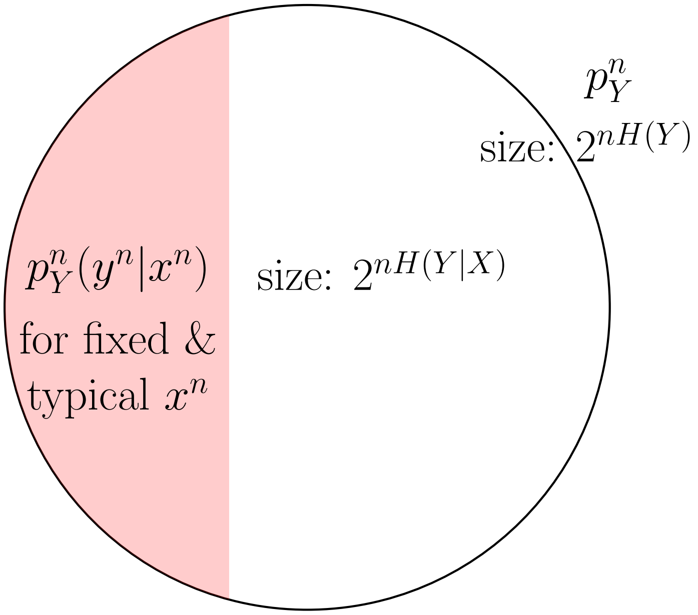
Applying the Markov's inequality for $a = 2\epsilon$, we have:

$$\Pr_{m}[\Pr_{N^n}[\text{error}|m] \geq 2\epsilon] \leq \Pr_{m}\left[\frac{\mathop{\mathbb{E}}_{N^n}([\text{error}|m])}{2\epsilon}\right] \leq \frac{1}{2}$$

Let's say the messages with $\Pr_{N^n}[\text{error}|m] \geq 2\epsilon$ are bad messages, then based on Markov's inequality, at most half of the messages are bad. But because the number of messages is exponential the number of channels, so it's no big deal to get rid of half of the bad messages ("expurgation").

The reduced codebook now will have at least half the size of the original codebook. For all the message $m$ in $C_{reduced}$, we have $\Pr_{N^n}[\text{error}|m] \leq 2\epsilon$

### 8.1.1 Proof intuition of the theorem



Bob receives strings from $p_Y^{\otimes n}$, i.e. $p_Y^n$ for classical case. The typical set of those has the size $2^{nH(Y)}$.

For any given message analysis, i.e. fixed typical $X^n$, then over a small subset we have the distribution $p_Y^n(y^n|x^n)$. The size of that subset is $2^{nH(Y|X)}$.

Why is that? For frequency typical, the number of $x$ appear in $x^n$ is approximately equal to $np_x(x)$. Suppose they are equal. For a string $x^n$, we have:

$$p_Y^n(y^n|x^n) = p(y_1|x_1)p(y_2|x_2)...p(y_n|x_n)$$

We expect $y^n$ to have:

$$\begin{aligned} np_x(x_1) &\text{ positions in the typical subspace } T^n_{p(\cdot|x_1),\delta} \\ np_x(x_2) &\text{ positions in the typical subspace } T^n_{p(\cdot|x_2),\delta} \end{aligned}$$

Then we can group all the $p$ with same $x$ together and have (but neglecting the $\delta$ stuffs just to provide intuitions):

$$\begin{aligned} p_Y^n(y^n|x^n) &= \prod_x \exp[-np_x(x)H(p(\cdot|x))] \\ &= \exp\left[-n\sum_x p_x(x)H(p(\cdot|x))\right] \\ &= \exp(-nH(Y|X)) \end{aligned}$$

If every string in subset has that probability, then the size of strings in the subset is roughly $2^{nH(Y|X)}$.

# 8.2 Converse of the Noisy Coding Theorem

## 8.2.1 Properties of entropy

- If $X$ is deterministic (for quantum, $\rho$ is pure), then $H(X) = 0$.
- If $Y$ is completely determined by $X$, i.e $Y = f(X)$, then $H(Y|X) = 0$.
- If $X$ and $Y$ are independent, i.e $p(X, Y) = p_x(X)p_y(Y)$ (for quantum, $\rho_{xy} = \rho_x \otimes \rho_y$), then $I(X : Y) = 0$.
- Conditional mutual information (CMI): If $X - Z - Y$ is a Markov chain, i.e $p(x, y, z) = p_Z(z)p(x|z)p(y|z)$, then $I(X : Y|Z) = 0$.

## 8.2.2 Properties of CMI

- Chain rule: $I(X : YZ) = I(X : Y) + I(X : Z|Y)$

  Proof: If we denote $H(\alpha)$ as $\alpha$ where $\alpha$ could be single or joint distribution and could also be conditional. Then we expand:

$$\begin{aligned}
\text{LHS} &= X - X|YZ = X - XYZ + YZ \\
\text{RHS} &= (X - X|Y) + (X|Y + Z|Y - XZ|Y) \\
&= (X - XY + Y) + (XY - Y + YZ - Y - XYZ + Y) \\
&= X + YZ - XYZ
\end{aligned}$$

  LHS and RHS are equal, thus the chain rule is proved.

- Generalized chain rule:

$$I(X : Y_1...Y_n) = I(X : Y_1) + I(X : Y_2|Y_1) + ... + I(X : Y_n|Y_1...Y_{n-1})$$

- Data processing inequality: If $X - Z - Y$ is a Markov chain, then as we move along the Markov chain, that should only degrade the mutual information, i.e $I(X : Z) \geq I(X : Y)$

Proof: Applying the chain rule above, we have:

$$\begin{aligned} I(X:Z) &= I(X:YZ) - I(X:Y|Z) \\ I(X:Y) &= I(X:YZ) - I(X:Z|Y) \end{aligned}$$

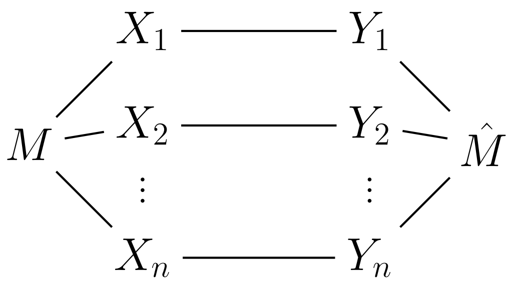
Take the difference on both sides of two equations:

$$\begin{aligned} I(X:Z) - I(X:Y) &= -I(X:Y|Z) + I(X:Z|Y) \\ &= I(X:Z|Y) \geq 0 \end{aligned}$$

In the second line, we used the property of Markov chain $I(X:Y|Z) = 0$. Thus $I(X:Z) \geq I(X:Y)$.

All of the above properties are true quantumly as well.

### 8.2.3 Converse of the Noisy Coding Theorem



If the noisy coding theorem says that we can send $nR$ bits and $R$ can get right up to the mutual information, the converse theorem says that we cannot do much better than that.

Consider a most general possible coding scheme: Alice sends message $M$, encodes it and inputs to the channels $X^n$. The input channels are mapped to output channels $Y^n$. Bob gets the outputs and decodes $\hat{M}$. i.e Markov chain $M - X^n - Y^n - \hat{M}$. We assume $M$ is uniformly distributed in $\{0,1\}^{nR}$, then $H(M) = nR$. Note that here we choose the uniform distribution here just for simplicity, and the theorem should apply to all possible distributions.

**Fano's inequality**

Obviously, the conditional entropy for $M$ based on $\hat{M}$ is small because for most cases they're equal, and similarly the mutual information between them is high. Quantitatively, we have the Fano's inequality says:

$$\begin{aligned} H(M|\hat{M}) &\le \epsilon nR + 1 \\ \to I(M:\hat{M}) = H(M) - H(M|\hat{M}) &\ge (1-\epsilon)nR - 1 \end{aligned}$$

Proof:

If the alphabet has size $d$, which in our real applications it's $nR$. And suppose that the probability of one element $p(m) \ge 1-\epsilon$, which corresponds to $M = \hat{M}$ in the above scenario. Then the entropy $H(p) \le 1 + \epsilon \log d$.

We name the $p(m) = 1-\delta, \delta \le \epsilon$. Rewrite the distribution $p = (1-\delta)1_m + \delta q$, where $q$ is another distribution which satisfies $q(m) = 0$. Then, the entropy can be rewritten as a sum of entropy of mixing being $m$ or not being $m$, and the entropy of the rest components:

$$\begin{aligned} H(p) &= -(1-\delta)\log(1-\delta) - \sum_x \delta q(x) \log \delta q(x) \\ &= -(1-\delta)\log(1-\delta) - \delta \log \delta - \delta \sum_x q(x) \log q(x) \\ &= H_2(\delta) + \delta H(q) \\ &\le 1 + \delta \log d \end{aligned}$$

Fannes' inequality (generalized version of Fano's inequality): If $p, q$ are distributions on alphabet of size $d$, then

$$\begin{aligned} |H(p) - H(q)| &\le H_2(\epsilon) + \epsilon \log d \\ \epsilon &= \frac{1}{2}||p-q||_1 \end{aligned}$$

Where in the quantum version of we just replace $H$ by $S$ and $p, q$ by the density matrix.

## Proof of converse theorem

Here we want to relate the above inequality to channels to work with Shannon's theorem:

$$(1-\epsilon)nR - 1 \underbrace{\le}_{\text{Fano's inequality}} I(M:\hat{M}) \overbrace{\le}^{\substack{\text{data processing worsen information} \\ \text{in Markov chain } M-X-Y-\hat{M}}} I(X^n:Y^n) \underbrace{\le}_{\star} \sum_{j=1}^{n} I(X_j:Y_j) \le nC \tag{8.1}$$

$$\to R \le \frac{C}{1-\epsilon}$$

The second inequality results from data processing: in the Markov's chain, the mutual information of two ends is less than or equal to the mutual information of the middles.

The last inequality: the mutual information of each input-output channels pair is at most $C$ (the mutual information obtained by maximizing over all the inputs).

The third inequality is a little bit unique because basically all other properties we mentioned in this section can be naturally generalized to quantum cases but this one not[1]. The major difference happens when the input has quantum entanglement. The reason we will mention the quantum capacity theorem in CQ channels in Sec. 7.2.1 is specifically to avoid such things to happen.

Now we try to prove (★) in classical regime. The mutual information is $I(X^n : Y^n) = H(Y^n) - H(Y^n|X^n)$. Because there is no correlation between different pairs of input-output channels, using chain rule, we have:

$$\begin{aligned} H(Y^n|X^n) &= \sum_{j=1}^{n} H(Y_j|X^n Y_1 ... Y_{j-1}) \\ &= \sum_{j=1}^{n} H(Y_j|X_j), \end{aligned}$$

where the last equality is based on the fact that the Markov chain only connects directly related pairs, so once condition on $X_j$, the $Y_j$ becomes conditionally independent on everything else. One can imagine that this fails quantumly when different $X_j$ are entangled and therefore can all contribute to $Y_j$.

The entropy of the sum is less than sum of the entropies of the part, i.e., the sub-additivity of entropy, so we have:

$$H(Y^n) \leq \sum_{j=1}^{n} H(Y_j)$$

[1] The conditional entropy, however, is also different in quantum since it can go to negative and therefore being equal to 0 does not have unique properties. The CMI and the corresponding Markov chain state can be generalized to quantum Markov states which we will revisit later, but in short the chain rule holds in quantum cases.

Thus,

$$\begin{aligned} I(X^n : Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &\leq \sum_{j=1}^{n} H(Y_j) - \sum_{j=1}^{n} H(Y_j|X_j) \\ &\leq \sum_{j=1}^{n} I(X_j : Y_j) \end{aligned}$$

## 8.3 Quantum Capacity Theorem

Idea: Find achievability via Packing Lemma

Example: Suppose that Alice has a menu of pure states as output $|0\rangle, |1\rangle, |+\rangle, |-\rangle$ to send

- Can send 1 classical bit ($0 \to |0\rangle$ and $1 \to |1\rangle$) or ($0 \to |+\rangle$ and $1 \to |-\rangle$)
- Can send 2 classical bits ($00 \to |0\rangle$, $01 \to |+\rangle$, $10 \to |-\rangle$ and $11 \to |1\rangle$)

Can Bob extract two classical bits from one quantum bit? No.

If $Q$ is the quantum system, then we have:

$$I(M : \hat{M}) \leq I(M : Q) \leq \log(\text{dim}Q) = 1$$

Therefore, Bob can extract at most one classical bit. So Alice should choose a distinguishable subset instead.

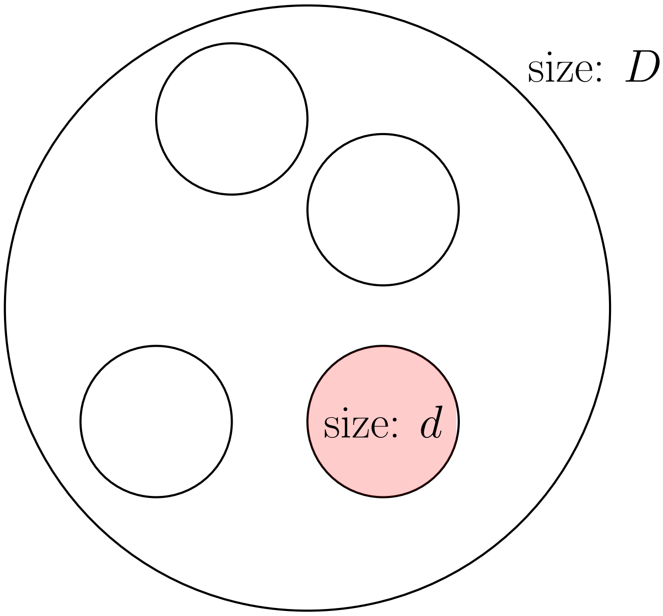
### 8.3.1 Packing Lemma

Given $\{\rho(x), \sigma(x)\}_{x \in X}$ with probability distribution $\rho(x)$ and signal state $\sigma(x)$. And $\sigma = \sum_x \rho(x)\sigma(x)$ is the average state.

Suppose there exists a projector $\Pi$ and family of projectors $\{\Pi_x\}_{x \in X}$ such that if $d$ and $D$ are dimensions of subspace and space then:

- $\text{Tr}[\Pi \sigma_x] \geq 1 - \epsilon$ for all $x$
- $\text{Tr}[\Pi_x \sigma_x] \geq 1 - \epsilon$ for all $x$
- $\text{Tr}[\Pi_x] \leq d$

- $\Pi\sigma\Pi \leq \frac{\Pi}{D}$



Choose the codebook $C = \{C_1, ..., C_M\} \sim p^n(n$ the number of letters in each code word), then there exists a positive operator-valued measure (POVM) $\{\Lambda_m\}$ such that averaging over the codebooks, averaging over messages in codebook, the probability to get the right outcome when measuring the state $\sigma_{C_m}$ is close to 1:

$$\mathop{\mathbb{E}}_{C} \mathop{\mathbb{E}}_{m\in[M]} \mathrm{Tr}(\Lambda_m \sigma_{C_m}) \geq 1 - 2\epsilon - 4\sqrt{\epsilon} - 4M\frac{d}{D}$$

Here whenever $M$ the total amount of message in codebook is less than the order of $D/d$ can give a small enough error probability. Corresponds to fill size $d \geq \mathrm{Tr}(\Pi_x)$ sphere inside of size $D(1-\epsilon) \geq \mathrm{Tr}(\Pi)$, like the Gaussian Noise channel in Sec. 7.1.1.

### 8.3.2 Application to channel coding

The HSW theorem (7.1) says the capacity of a noisy quantum channel is the maximal mutual information between input $X$ and output $Q$, maximizing over input distribution $p$

$$C(N) = \max_p I(X:Q)$$

Choosing the codebook $C_i$ from $p^n_T = p^n|T^n_{p,\delta}$ where $T^n_{p,\delta}$ is the frequncy-typical set.

Then, the corresponding string $x^n$ and state $\rho_{x^n}$ are:

$$x^n(i) = C_i$$
$$\rho_{x^n} = \rho_{x_1} \otimes \rho_{x_2} \otimes ... \otimes \rho_{x_n}$$

We need to show this choice satisfies all the conditions of Packing Lemma. The average state:

$$\sigma = \mathbb{E}(\rho_{x^n}) = \sum_{x^n} p_T^n(x^n)\rho_{x^n} \approx \sum_{x^n} p^n(x^n)\rho_{x^n} = \bar{\rho}^{\otimes n} \text{ for } \bar{\rho} = \sum_x p(x)\rho_x$$

And let's take the total projector to be projecting into this average state: $\Pi = \Pi^n_{\bar{\rho},\delta}$. Does this projector satisfy the four conditions in packing Lemma?

- Condition 1:

$$\mathop{\mathbb{E}}_{x^n} \operatorname{Tr}[\Pi\rho_{x^n}] \geq \operatorname{Tr}[\Pi\bar{\rho}^{\otimes n}] - \epsilon \geq 1 - 2\epsilon$$

Therefore, the best 1/2 of $x^n \in X^n$ have $\operatorname{Tr}[\Pi\rho_{x^n}] \geq 1 - 4\epsilon$

Denote $\Pi_{X^n}$ as the conditionally typical projector, it is calculated as follow:

$$\Pi_{X^n} = \bigotimes_{x \in X} \Pi^{\#x}_{\rho_x,\delta}$$

This product is permuted according to $x^n$ and $\#x$ is the count of occurrences of $x$ in $x^n$.

- Condition 2:

$$\operatorname{Tr}[\Pi_{X^n}\rho_{X^n}] = \prod_{x \in X} \operatorname{Tr}[\rho_x^{\otimes \#x}\Pi^{\#x}_{\rho_x,\delta}] \geq 1 - |X|\epsilon$$

- Condition 3:

$$\begin{aligned}
\operatorname{Tr}[\Pi_{X^n}] &= \prod_{x \in X} \operatorname{Tr}[\Pi^{\#x}_{\rho_x,\delta}] \\
&\leq \prod_{x \in X} \operatorname{Tr}[\Pi^{n(p(x)+\delta)}_{\rho_x,\delta}] \text{ because of the freq typicality} \\
&\leq \prod_{x \in X} \exp(n(p(x)+\delta)(S(\rho_x)+\delta)) \\
&\leq \exp(n(S(Q|X)+\delta'))
\end{aligned}$$

- Condition 4: For $p_T^n \leq (1-\epsilon)^{-1}p^n$, we have the average of $\rho_{X^n}$ is

$$E(\rho_{X^n}) \leq (1-\epsilon)^{-1} \sum_{X^n} p^n(X^n)\rho_{X^n} = (1-\epsilon)^{-1}\bar{\rho}^{\otimes n}$$

Now we need to calculate

$$\Pi\bar{\rho}^{\otimes n}\Pi \leq 2^{-n(S(\rho)-\delta)}$$
$$\rightarrow \Pi E(\rho_{X^n})\Pi \leq (1-\epsilon)^{-1}2^{-n(S(\rho)-\delta)}$$

Here $D \approx 2^{nS(Q)}$, $d \approx 2^{nS(Q|X)}$, thus we can take $M \approx 2^{nI(X:Q)}$.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 9: Sep 29, 2020

*Lecturer: Aram Harrow* *Scribe: John Martyn, Thao Dinh*

## 9.1 Details on the Packing Lemma

Regarding the packing lemma, one might be curious as to why we cannot simply use the POVM $\{\Pi_m\}$ in place of $\{\Lambda_m\}$. We can illustrate that this is not correct with a classical instance of the packing lemma. We also show how to correct this misconception and properly define $\{\Lambda_m\}$.

### 9.1.1 Classical Case

Look at the simple example in which the codebook is $C = \{1, 2\}$, and the projectors are $\Pi_1 = \text{diag}(1, 1, 1, 1, 0, 0, 0)$ and $\Pi_2 = \text{diag}(0, 0, 0, 1, 1, 1, 1)$. We will take as our signal states $\sigma_1 = \Pi_1/4$ and $\sigma_2 = \Pi_2/4$. As all the operators are diagonal, this is essentially a classical problem.

Anyhow, observe that $\Pi_1 + \Pi_2 = \text{diag}(1, 1, 1, 2, 1, 1, 1) \neq I$, so $\{\Pi_1, \Pi_2\}$ is not a valid POVM. Hence, we cannot simply use $\{\Pi_1, \Pi_2\}$ in place of $\{\Lambda_1, \Lambda_2\}$

However, in this scenario, we can instead use as a POVM $\Lambda_1 = \text{diag}(1, 1, 1, \frac{1}{2}, 0, 0, 0)$ and $\Lambda_2 = \text{diag}(0, 0, 0, \frac{1}{2}, 1, 1, 1)$. In general, when dealing with classcal instances of the packing lemma, we can obtain valid POVM $\{\Lambda_m\}$ as follows:

$$\Pi_{\text{total}} = \sum_m \Pi_m$$
$$\Lambda_m = \Pi_{\text{total}}^{-1} \Pi_m.$$

A quick calculation indices that this yields the above expression for $\{\Lambda_1, \Lambda_2\}$.

### 9.1.2 Quantum Case

In the quantum case of the packing lemma, we again cannot simply set $\Lambda_m = \Pi_m$. Instead, we can use the following prescription to construct $\Lambda_m$, which is analogous to the procedure employed above. Let $P_m = \Pi\Pi_m\Pi$ and $P_{\text{total}} = \sum_m P_m$. Note that

$P_m \geq 0$ because it is constructed as a symmetric product of projection operators. We then define $\Lambda_m = P_{\text{total}}^{-1/2} P_m P_{\text{total}}^{-1/2}$. If it is the case that $P_{\text{total}}$ is not full rank, introduce an additinal projector $\Lambda_{\text{fail}}$ such that $\sum_m \Lambda_m = I$. $\{\Lambda_m\}$ is then a valid POVM.

## 9.2 Packing Lemma Proof

We can use the above prescription to prove the packing lemma. To begin, let's look at the probability of incurring an error on message $m$, given codebook C:

$$p_{\text{err}}(m|C) = 1 - \text{tr}(\Lambda_m \sigma_m) = \text{tr}\Big((I - \Lambda_m)\sigma_m\Big)$$

To analyze this expression, we will employ the Hayashi-Nagaoka lemma: Given $S$ and $T$, such that $0 \leq S \leq I$ and $T \geq 0$,

$$I - (S+T)^{-1/2} S (S+T)^{-1/2} \leq 2(I - S) + 4T.$$

Set $T = \sum_{m \neq m'} P_{m'} = P_{\text{total}} - P_m$, and $S = P_m$. These obey $0 \leq S \leq I$ and $T \geq 0$, so we can apply the Hayashi-Nagaoka lemma:

$$I - \Lambda_m = I - P_{\text{total}}^{-1/2} P_m P_{\text{total}}^{-1/2} =$$
$$I - (S+T)^{-1/2} S (S+T)^{-1/2} \leq 2(I - P_m) + 4 \sum_{m' \neq m} P_{m'}.$$

Using this result to evaluate $p_{\text{err}}(m|C)$, we have

$$p_{\text{err}}(m|C) \leq 2(1 - \text{tr}(P_m \sigma_m)) + 4 \sum_{m' \neq m} \text{tr}(P_{m'} \sigma_m)$$

Then, making use of the conditions assumed in the packing lemma, we can establish the bound

$$\text{tr}(P_m c_m) = \text{tr}(\Pi \Pi_m \Pi \sigma_m) = \text{tr}(\Pi_m \Pi \sigma_m \Pi) \geq$$
$$\text{tr}(\Pi_m \sigma_m) - ||\sigma_m - \Pi \sigma_m \Pi||_2 \geq 1 - \epsilon - 2\sqrt{\epsilon},$$

where we obtain the $\sqrt{\epsilon}$ as a result of the bound on gentle measurement proven in problem set 3. Next, we input the above bound into the expression for $p_{\text{err}}(m|C)$ and average this probability over messages $m$ and codebooks $C$:

$$\mathop{\mathbb{E}}_C \mathop{\mathbb{E}}_m p_{\text{err}}(m|C) \leq 2(\epsilon + 2\sqrt{\epsilon}) + 4 \mathop{\mathbb{E}}_m \frac{1}{M} \sum_{m' \neq m} \text{tr}(P_{m'} \sigma_m).$$

Noting that $\mathop{\mathbb{E}}_C \sigma_m = \mathop{\mathbb{E}}_C \sigma_{C_m} = \sum_x p(x) \sigma_x = \sigma$, we can write the second term as

$$\frac{1}{M} \sum_{m \neq m'} \text{tr}(\Pi \Pi_{m'} \Pi) \mathop{\mathbb{E}}_C \sigma_m = \sum_{m \neq 1} \text{tr}(\Pi_{m'} \Pi \sigma \Pi) \leq (M-1) \text{tr}\Big(\Pi_{m'} \frac{I}{D}\Big) \leq M \frac{d}{D},$$

where we have employed the inequalities assumed in the packing lemma. Combining all the terms and inequalities above, we have

$$p_{\mathrm{err}}(m|C) = 1 - \mathrm{tr}(\Lambda_m \sigma_m) \leq 2\epsilon + 4\sqrt{\epsilon} + 4M\frac{d}{D} \quad \Rightarrow$$
$$\mathrm{tr}(\Lambda_m \sigma_m) \geq 1 - 2\epsilon - 4\sqrt{\epsilon} - 4M\frac{d}{D}.$$

This is the claim of the packing lemma, which is now proven.

## 9.3 Aside: Pretty Good Measurement

Imagine that given a state $\sigma = \sum_x p(x)\sigma_x$, we wish to distinguish between the states $\sigma_x$. We can do this decently well with the "pretty good measurement" which is defined by the POVM $M_x = \sigma^{-1/2} p(x)\sigma_x \sigma^{-1/2}$. The Barnum-Knill theorem proves that the pretty good measurement can distinguish between the states $\sigma_x$ with error probability

$$p_{\mathrm{err}}(\text{Pretty Good Measurement}) \leq 2p_{\mathrm{err}}(\text{Optimal Measurement}).$$

So in general, the "pretty good measurement" achieves an error probability that is comparable to the optimal error probability.

The pretty good measurement can be thought of as reversing the action of the channel $\mathcal{N} : x \to \sigma_x$, and applying this reversal to the state $\rho = \sum_x p(x)|x\rangle\langle x|$. In particular, if $\mathcal{N}$ has Kraus operators $\{E_k\}$, then the reversal of this channel, which we call the recovery channel, has Kraus operators $F_k = \rho^{1/2} E_k^\dagger \rho^{-1/2}$. This generalizes the "pretty good measurement" to a more general construction known as the "Petz recovery map".

## 9.4 Sequential Coding

In sequential decoding, one decodes a message by enumerating through the set of all possible message sequences. Specifically, we are given a state $\sigma_x$, and perform on it the set of measurements $\{\Pi, I - \Pi\}$, $\{\Pi_{c_1}, I - \Pi_{c_1}\}$, ..., $\{\Pi_{c_m}, I - \Pi_{c_m}\}$. These measurements dictate whether we fail or continue in the sequential coding procedure as follows

$$\Pi \to \text{continue}, \quad I - \Pi \to \text{fail}$$
$$\Pi_{c_m} \to \text{stop, output m}, \quad I - \Pi_{c_m} =: \hat{\Pi}_{c_m} \to \text{continue}$$

The probability that this procedure fails to output $m$ is

$$p_{\mathrm{err}}(m) = 1 - p_{\mathrm{success}} = 1 - \mathrm{tr}\Big(\Pi_{C_m}\hat{\Pi}_{c_{m-1}}...\hat{\Pi}_{c_1}\Pi\sigma_{c_m}\Pi\hat{\Pi}_{c_1}\hat{\Pi}_{c_{m-1}}\Pi_{C_m}\Big).$$

To analyze this expression, which we will do in the future, we will make use of the non-commutative union bound:

$$\omega \geq 0, \quad \mathrm{tr}(\omega) \leq 1, \quad P_1, ..., P_L = \text{set of projectors} \quad \Rightarrow$$
$$\mathrm{tr}(\omega) - \mathrm{tr}(P_L ... P_1 \omega P_1 ... P_L) \leq \sqrt{2 \sum_i \mathrm{tr}(\hat{P}_i \omega)}, \quad \hat{P}_i = I - P_i.$$

We will prove this relation next class. For now, we can observe that it is not at all obvious. Imagining that $\omega$ is a density matrix, the above quantity on the LHS will measure the difference between the density matrix, and the density matrix after a set of $L$ projective measurements are applied to it. In general, applying a set of projective measurements can change the state drastically. For instance, imagine states

$$|\phi_j\rangle = \cos\left(\frac{\pi}{2}\frac{j}{L}\right)|0\rangle + \sin\left(\frac{\pi}{2}\frac{j}{L}\right)|1\rangle, \quad j = 1, ..., L,$$

to which we apply projectors

$$P_j = |\phi_j\rangle\langle\phi_j|.$$

With this setup, we have $\langle\phi_j|P_{j+1}|\phi_j\rangle = |\langle\phi_j|\phi_{j+1}\rangle|^2 = \cos^2(\frac{\pi}{2}\frac{1}{L}) = 1 - O(L^{-2})$, which indicates that applying the measurement $P_{j+1}$ to $|\phi_j\rangle$, transitions one to state $|\phi_{j+1}\rangle$ with high probability. Therefore, one can begin in the state $|\phi_0\rangle \approx |0\rangle$ and end in $|\phi_L\rangle \approx |1\rangle$ with probability $1 - O(\frac{1}{L})$. These states are very different from each other (nearly orthogonal!), so it can be tricky to place a bound on $\mathrm{tr}(\omega) - \mathrm{tr}(P_L ... P_1 \omega P_1 ... P_L)$. We will discuss this further in the next class.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 10: October 1, 2020

*Lecturer: Aram Harrow* *Scribe: Matthew Kearney, John Martyn*

## 10.1 Non-Commutative Union Bound

To begin, recall the statement of the non-commutative union bound:

$$\omega \geq 0, \quad \mathrm{tr}(\omega) \leq 1, \quad P_1, ..., P_L = \text{set of projectors} \quad \Rightarrow$$
$$\mathrm{tr}(\omega) - \mathrm{tr}(P_L ... P_1 \omega P_1 ... P_L) \leq 2\sqrt{\sum_i \mathrm{tr}(\hat{P}_i \omega)}, \quad \hat{P}_i = I - P_i.$$

We will now prove this bound. We will first examine the case where $\omega$ is a pure state written as

$$\omega = |\psi\rangle \langle\psi| , \quad \| |\psi\rangle \| \leq 1$$

We would like to show that

$$\| |\psi\rangle \|^2 - \|P_L ... P_1 |\psi\rangle \|^2 \leq 2\sqrt{\sum_i \|\hat{P}_i |\psi\rangle \|^2}.$$

We note that since $P_L$ and $\hat{P}_L$ are orthogonal operators that sum to $I$ we can write $|\psi\rangle$ as

$$|\psi\rangle = P_L |\psi\rangle + \hat{P}_L |\psi\rangle .$$

We will now use a proof by induction on $L$ with the inductive assumption that

$$\| |\psi\rangle - P_{L-1} ... P_1 |\psi\rangle \|^2 \leq \sum_{i=1}^{L-1} \|\hat{P}_i |\psi\rangle \|^2.$$

To begin we have

$$|\psi\rangle - P_L ... P_1 |\psi\rangle = \hat{P}_L |\psi\rangle + P_L(|\psi\rangle - P_{L-1} ... P_1 |\psi\rangle).$$

Since $P_L$ and $\hat{P}_L$ are orthogonal operators we can then use the Pythagorean theorem

$$\| \left|\psi\right\rangle - P_L ... P_1 \left|\psi\right\rangle \|^2 = \|\hat{P}_L \left|\psi\right\rangle \|^2 + \|P_L(\left|\psi\right\rangle - P_{L-1}...P_1 \left|\psi\right\rangle)\|^2.$$

Since projection operators do not increase the norm we have

$$\| \left|\psi\right\rangle - P_L ... P_1 \left|\psi\right\rangle \|^2 \leq \|\hat{P}_L \left|\psi\right\rangle \|^2 + \| \left|\psi\right\rangle - P_{L-1}...P_1 \left|\psi\right\rangle \|^2.$$

We can then use our inductive assumption to get

$$\| \left|\psi\right\rangle - P_L ... P_1 \left|\psi\right\rangle \|^2 \leq \sum_{i=1}^{L} \|\hat{P}_i \left|\psi\right\rangle \|^2 \quad \Rightarrow$$

$$\| \left|\psi\right\rangle - P_L ... P_1 \left|\psi\right\rangle \| \leq \sqrt{\sum_{i=1}^{L} \|\hat{P}_i \left|\psi\right\rangle \|^2}.$$

On the other hand, by the triangle inequality we get

$$\| \left|\psi\right\rangle \| - \|P_L ... P_1 \left|\psi\right\rangle \| \leq \sqrt{\sum_{i=1}^{L} \|\hat{P}_i \left|\psi\right\rangle \|^2}.$$

Let $A = \| \left|\psi\right\rangle \|$ and $B = \|P_L ... P_1 \left|\psi\right\rangle \|$. Since $A, B \leq 1$ we have

$$A^2 - B^2 = (A-B)(A+B) \leq 2(A-B) \leq 2\sqrt{\sum_{i=1}^{L} \|\hat{P}_i \left|\psi\right\rangle \|^2}.$$

This proves the statement of the the theorem for the pure state case. We now discuss the case of mixed states. We note that the left hand side of the inequality is linear in omega. Therefore if $\omega = \sum\limits_i p_i \psi_i$, this gives us

$$\operatorname{tr}(\omega) - \operatorname{tr}(P_L ... P_1 \omega P_1 ... P_L) = \sum_i p_i (\operatorname{tr}(\psi_i) - \operatorname{tr}(P_L ... P_1 \psi_i P_1 ... P_L)).$$

However, the right hand side of the inequality consists of the square root of a linear function of $\omega$. This means the right hand side is concave in $\omega$. This gives us the following property

$$2\sqrt{\sum_i \operatorname{tr}(\hat{P}_i \omega)} = 2\sqrt{\sum_i \sum_j p_j \operatorname{tr}(\hat{P}_i \psi_j)} \geq 2 \sum_j p_j \sqrt{\sum_i \operatorname{tr}(\hat{P}_i \psi_j)}.$$

Putting these facts together we have

$$\begin{aligned}\mathrm{tr}(\omega) - \mathrm{tr}(P_L ... P_1 \omega P_1 ... P_L) &= \sum_i p_i(\mathrm{tr}(\psi_i) - \mathrm{tr}(P_L ... P_1 \psi_i P_1 ... P_L)) \\ &\leq 2 \sum_j p_j \sqrt{\sum_i \mathrm{tr}(\hat{P}_i \psi_j)} \\ &\leq 2 \sqrt{\sum_i \mathrm{tr}(\hat{P}_i \omega)}.\end{aligned}$$

This proves the statement of the theorem for mixed states.

## 10.2 Proving HSW Theorem with Non-Commutative Union Bound

Recall that the failure probability of our sequential decoding scheme is given by

$$p_{\mathrm{err}}(m) = 1 - p_{\mathrm{success}} = 1 - \mathrm{tr}\Big(\Pi_{c_m} \hat{\Pi}_{c_{m-1}} ... \hat{\Pi}_{c_1} \Pi \sigma_{c_m} \Pi \hat{\Pi}_{c_1} \hat{\Pi}_{c_{m-1}} \Pi_{c_m}\Big).$$

Remember from the hypothesis of the packing lemma that

$$\mathrm{tr} \Pi \sigma_{c_m} \Pi \geq 1 - \epsilon.$$

Putting these together we have

$$p_{\mathrm{err}}(m) \leq \epsilon + \mathrm{tr} \Pi \sigma_{c_m} \Pi - \mathrm{tr}(\Pi_{c_m} \hat{\Pi}_{c_{m-1}} ... \hat{\Pi}_{c_1} \Pi \sigma_{c_m} \Pi \hat{\Pi}_{c_1} \hat{\Pi}_{c_{m-1}} \Pi_{c_m}).$$

Using the non-commutative union bound to $\Pi \sigma_{c_m} \Pi$, we get

$$p_{\mathrm{err}}(m) \leq \epsilon + 2 \sqrt{\mathrm{tr}((\hat{\Pi}_{c_m} + \Pi_{c_{m-1}} + ... + \Pi_{c_1}) \Pi \sigma_{c_m} \Pi)}$$

Taking the expectation of this quantity over the message and codebook, we establish

$$\mathbb{E}_C \mathbb{E}_m p_{\mathrm{err}}(m|C) \leq \epsilon + 2 \mathbb{E}_{m,C} \sqrt{\mathrm{tr}((\hat{\Pi}_{c_m} + \Pi_{c_{m-1}} + ... + \Pi_{c_1}) \Pi \sigma_{c_m} \Pi)}.$$

Once again, by the concavity of the square root (i.e. applying Jensen's inequality), we have that

$$\mathbb{E}_C\mathbb{E}_m p_{\text{err}}(m|C) \le \epsilon + 2\sqrt{\mathbb{E}_{m,C}\text{tr}((\hat{\Pi}_{c_m} + \Pi_{c_{m-1}} + ... + \Pi_{c_1})\Pi\sigma_{c_m}\Pi)}$$

We already showed in section 9.2 that

$$\mathbb{E}_{m,C}\text{tr}(\hat{\Pi}_{c_m}\Pi\sigma_{c_m}\Pi) \le \epsilon + 2\sqrt{\epsilon}$$
$$\mathbb{E}_{m,C}\sum_{m\neq m'}\text{tr}(\Pi_{c'_m}\Pi\sigma_{c_m}\Pi) \le \frac{Md}{D}$$

Therefore this gives us

$$\mathbb{E}_{m,C}p_{\text{err}}(m|C) \le \epsilon + 2\sqrt{\epsilon + 2\sqrt{\epsilon} + Md/D}.$$

# Hypothesis Testing

We would like to distinguish $\rho^{\otimes n}$ from $\sigma^{\otimes n}$. Specifically, we want a measurement $M$ such that

$$\text{tr}(\rho^{\otimes n}M) \ge \alpha, \quad \alpha \in (0,1)$$
$$\text{tr}(\sigma^{\otimes n}M) \sim 2^{-nR}$$

We will prove Stein's Lemma, which states the optimal $R = D(\rho\|\sigma) = \text{tr}(\rho(\log\rho - \log\sigma))$. The optimal $M$ is the projector onto $[\alpha^{-1}\rho^{\otimes n} - 2^{nR}\sigma^{\otimes n} \ge 0]$ which is the projector onto the non-negative eigenspace of the given quantity.

We have shown as an exercise that, classically, the best $M$ to distinguish distributions $p^n$ and $q^n$ is given by $M$ being a projector onto $T^n_{p,\delta}$. Specifically,

$$p^n(T^n_{p,\delta}) \to 1 \text{ as } n \to \infty$$
$$q^n(T^n_{p,\delta}) \approx |T^n_{p,\delta}|q(1)^{np(1)}...q(d)^{np(d)} \approx 2^{-nD(p\|q)}$$

so we can distinguish the two stat fairly well, depending on the magnitude of $D(p\|q)$.

We will now explore the quantum version following the proof of Bjelakovic et al. Define $\rho$ and $\sigma$ as

$$\rho = \sum_x r_x |\alpha_x\rangle\langle\alpha_x| \quad \sigma = \sum_x s_x |\beta_x\rangle\langle\beta_x|$$

We define a new type of typical projector as

$$\Pi^n_{\rho\|\sigma,\delta} = \sum_{x^n : |\frac{1}{n}\sum_{i=1}^{n} \log s_{x_i} - \mathrm{tr}(\rho\log\sigma)| \leq \delta} \beta_{x^n}$$

$$\beta_{x^n} = \beta_{x_1} \otimes ... \otimes \beta_{x_n}$$

We note the following properties of this projector

$$\mathrm{tr}(\rho^{\otimes n}\Pi^n_{\rho\|\sigma,\delta}) \geq 1-\epsilon \tag{10.1}$$

$$[\Pi^n_{\rho\|\sigma,\delta}, \sigma^{\otimes n}] = 0 \tag{10.2}$$

$$2^{n\mathrm{tr}(\rho\log\sigma-\delta)}\Pi^n_{\rho\|\sigma,\delta} \leq \Pi^n_{\rho\|\sigma,\delta}\sigma^{\otimes n}\Pi^n_{\rho\|\sigma,\delta} \leq 2^{n\mathrm{tr}(\rho\log\sigma+\delta)}\Pi^n_{\rho\|\sigma,\delta} \tag{10.3}$$

## Achievability

We will first show that Stein's Lemma is achievable with $M = \Pi^n_{\rho\|\sigma,\delta}\Pi^n_{\rho,\delta}\Pi^n_{\rho\|\sigma,\delta}$. With this definition, we have

$$\begin{aligned}\mathrm{tr}(\rho^{\otimes n}\Pi^n_{\rho,\delta} - \rho^{\otimes n}M) &= \mathrm{tr}(\Pi^n_{\rho,\delta}(\rho^{\otimes n} - \Pi^n_{\rho,\delta}\Pi^n_{\rho\|\sigma,\delta}\rho^{\otimes n}\Pi^n_{\rho\|\sigma,\delta}) \\ &\leq \|\rho^{\otimes n} - \Pi^n_{\rho\|\sigma,\delta}\rho^{\otimes n}\Pi^n_{\rho\|\sigma,\delta}\|_1 \quad \Rightarrow \\ \mathrm{tr}(M\rho^{\otimes n}) &\geq \mathrm{tr}(\rho^{\otimes n}\Pi^n_{\rho,\delta}) - \|\rho^{\otimes n} - \Pi^n_{\rho\|\sigma,\delta}\rho^{\otimes n}\Pi^n_{\rho\|\sigma,\delta}\|_1.\end{aligned}$$

By the gentle measurement lemma we have that

$$\mathrm{tr}(M\rho^{\otimes n}) \geq 1-\epsilon-2\sqrt{\epsilon} \geq \alpha.$$

Now we look at how $M$ acts on $\sigma^{\otimes n}$:

$$\mathrm{tr}(M\sigma^{\otimes n}) = \mathrm{tr}(\Pi^n_{\rho,\delta}\Pi^n_{\rho\|\sigma,\delta}\sigma^{\otimes n}\Pi^n_{\rho\|\sigma,\delta}).$$

Using equation 10.3 this gives us

$$\mathrm{tr}(M\sigma^{\otimes n}) \leq \mathrm{tr}(\Pi^n_{\rho,\delta})2^{n\mathrm{tr}(\rho\log\sigma+\delta)} \leq 2^{n(S(\rho)+\delta\mathrm{tr}(\rho\log\sigma)+\delta)} = 2^{-n(D(\rho\|\sigma)-2\delta)},$$

and so we have proven achievability.

## Converse

Suppose $\operatorname{tr}(M\rho^{\otimes n}) \geq \alpha$. We will argue that $\operatorname{tr}(M\sigma^{\otimes n})$ is not too small. From 10.2 and 10.3 we have

$$\begin{aligned} \sigma^{\otimes n} &\geq \Pi^n_{\rho\|\sigma,\delta} 2^{n\operatorname{tr}(\rho\log\sigma - \delta)} \\ \operatorname{tr}(M\sigma^{\otimes n}) &\geq \operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta}) 2^{n\operatorname{tr}(\rho\log\sigma-\delta)}. \end{aligned}$$

To bound this, we will now show a bound for $\operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta})$. We note the following

$$\rho^{\otimes n}\Pi^n_{\rho,\delta} = \Pi^n_{\rho,\delta}\rho^{\otimes n}\Pi^n_{\rho,\delta} \leq 2^{(-n(s(\rho)-\delta))}\Pi^n_{\rho,\delta} \tag{10.4}$$

We will compute $\operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta})$.

$$\begin{aligned} \operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta}) &= \operatorname{tr}(\Pi^n_{\rho\|\sigma,\delta}M\Pi^n_{\rho\|\sigma,\delta}) \\ &\geq \operatorname{tr}(\Pi^n_{\rho\|\sigma,\delta}M\Pi^n_{\rho\|\sigma,\delta}\Pi^n_{\rho,\delta}) \end{aligned}$$

Using equation 10.4 we have

$$\operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta}) \geq \operatorname{tr}(\Pi^n_{\rho\|\sigma,\delta}M\Pi^n_{\rho\|\sigma,\delta}\Pi^n_{\rho,\delta}\rho^{\otimes n})2^{n(s(\rho)-\delta)}.$$

Let $B$ be the atypical part of $\rho^{\otimes n}$ ($\rho = A + B =$ typical + atypical).

$$\operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta}) \geq \operatorname{tr}(\Pi^n_{\rho\|\sigma,\delta}M\Pi^n_{\rho\|\sigma,\delta}(\rho^{\otimes n} - B))2^{n(s(\rho)-\delta)}.$$

Once again by gentle measurement we have

$$\operatorname{tr}(M\Pi^n_{\rho\|\sigma,\delta}) \geq (\alpha - 2\sqrt{\epsilon} - \epsilon)2^{n(S(\rho)-\delta))}.$$

This finally brings us to our conclusion that

$$\operatorname{tr}(M\sigma^{\otimes n}) \geq (\alpha - 2\sqrt{\epsilon} - \epsilon)2^{-n(D(\rho\|\sigma)+2\delta))},$$

and the proof of the converse is complete.

## Corollary: Monotonicity of $D(\rho\|\sigma)$ under Partial Trace

Given $\rho_{AB}$,$\sigma_{AB}$ there exists an $M$ such that $\mathrm{tr}(M\rho_A^{\otimes n}) \geq \alpha$ and $\mathrm{tr}(M\sigma_A^{\otimes n}) \approx 2^{-nD(\rho_A\|\sigma_A)}$. This means

$$\mathrm{tr}((M \otimes I_B)^{\otimes n}\rho_{AB}^{\otimes n}) = \mathrm{tr}(M\rho_A^{\otimes n}) \geq \alpha$$
$$\mathrm{tr}((M \otimes I_B)^{\otimes n}\sigma_{AB}^{\otimes n}) = \mathrm{tr}(M\sigma_A^{\otimes n}) \geq 2^{-nD(\rho_A\|\sigma_A)}.$$

Therefore

$$2^{-nD(\rho_A\|\sigma_A)} \geq 2^{-nD(\rho_{AB}\|\sigma_{AB})} \quad \Rightarrow$$
$$D(\rho_{AB}\|\sigma_{AB}) \geq D(\rho_A\|\sigma_A).$$

Evidently, $D(\|)$ is decreases under partial trace.

## Corollary: Strong Subadditivity

We can express the conditional mutual information as

$$I(A : C|B) = I(A : BC) - I(A : B) = D(\rho_{ABC}\|\rho_B \otimes \rho_{BC}) - D(\rho_{AB}\|\rho_A \otimes \rho_B).$$

If we let $\sigma_{ABC} = \rho_B \otimes \rho_{BC}$, then the second divergence is simply the first but with both systems traced over C. Thus, the monotonicity of $D(\rho\|\sigma)$ under partial trace gives us that

$$I(A : C|B) = I(A : BC) - I(A : B) \geq 0.$$

This is just strong subadditivity.

## Aside: Converse of Schumacher Compression

Recall equation 10.4

$$A = \rho^{\otimes n}\Pi_{\rho,\delta}^n = \Pi_{\rho,\delta}^n\rho^{\otimes n}\Pi_{\rho,\delta}^n \leq 2^{-n(s(\rho)-\delta)}\Pi_{\rho,\delta}^n$$

Let $\rho^{\otimes n} = A + B$ with $\mathrm{tr}(B) \leq \epsilon$. Then we have

$$\alpha \leq \mathrm{tr}(M\rho^{\otimes n}) = tr(MA) + \mathrm{tr}(BM) \leq \mathrm{tr}(AM) + \epsilon.$$

This gives us

$$\alpha - \epsilon \leq \mathrm{tr}(AM) \leq \mathrm{tr}(M\Pi^n_{\rho,\delta}) \exp(-n(S(\rho) - \delta)).$$

So finally we have that

$$\mathrm{tr}(M) \geq \mathrm{tr}(M\Pi^n_{\rho,\delta}) \geq (\alpha - \epsilon) \exp(n(S(\rho) - \delta)),$$

which is the converse of Schumacher compression.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 11: October 6, 2020

*Lecturer: Aram Harrow* *Scribe: Adam Bene Watts, Matthew Kearney*

In this section, we discuss some of the applications of relative entropy to show that it is a useful measure.

## 11.1 Application 1: Channel Coding

Consider a CQ channel $\{p(x), \sigma_x\}$, where message $x$ is sent with probability $p_x$ and $\sigma_x$ is the resulting signal state. Define $\sigma = \sum_x p(x)\sigma_x$, making $\sigma$ the average over the input states.

Recall that the relative entropy is given by

$$D(\sigma_x||\sigma) = \mathrm{tr}[\sigma_x(\log(\sigma_x) - \log(\sigma)] = -S(\sigma_x) - \mathrm{tr}[\sigma_x(\log(\sigma))]$$

If we then take the average over these relative entropies, we get the familiar Holevo $\chi$, which describes the difference between the entropy of the average state and the average of the entropies of each of the states.

$$\begin{aligned} \sum_x p(x) D(\sigma_x||\sigma) &= -\sum_x p(x) S(\sigma_x) - \mathrm{tr}\left[\sum_x p(x)\sigma_x \log(\sigma)\right] \\ &= S(\sigma) - \sum_x p(x) S(\sigma_x) = \chi. \end{aligned}$$

This is an interesting result and leads us to ask Why should the relative entropy have anything to do with the channel capacity?

We can think of this as saying that the ability of the ensemble to carry information is related to the how surprising each message $\sigma_x$ is compared to the average state $\sigma$, which is given by $D(\sigma_x||\sigma)$. To give a classical example, if we imagine that it rains 10% of the time and is sunny the other 90%, then the relative entropy between the state rainy and the average state will be low, meaning we are less surprised about it being sunny when the average state is our prior.

Along this line of thinking, we can imagine a hypothesis testing scenario where we try to distinguish a typical message $\sigma_{x^n} = \sigma_{x_1} \otimes \sigma_{x_2} \otimes ... \otimes \sigma_{x_n}$ from the average state

$\sigma^{\otimes n}$, which serves as our prior of the messages we will receive. Stein's Lemma tells us we mistakenly identify the message as the average state with probability $2^{-n\chi}$. This quantity is important for hypothesis testing realizations like sequential decoding where we may need to test against exponentially many possible states before testing against the correct state and we want to be very sure that we are not accepting the wrong messages.

While this discussion is suggestive of a strong link between hypothesis testing and channel coding, Ogawa and Nagaoka formalized this link by showing that you can prove the HSW theorem using hypothesis testing with carefully chosen states.

## 11.2 Application 2: Thermal States

Let $H$ be a Hamiltonian and define

$$\gamma_T = \frac{e^{-H/T}}{\mathrm{tr}[e^{-H/T}]} \qquad\qquad F(\rho) = E(\rho) - TS(\rho) = \mathrm{tr}[H\rho] - TS(\rho).$$

Where $F(\rho)$ is the free energy and the thermal state $\gamma_T$ is the state that minimizes free energy. Recall from PSET 4 that We derived an expression for a measure of how close a state's free energy is to the minimum free energy given by

$$\frac{D(\rho||\gamma_T)}{\ln(2)} = \frac{F(\rho) - F(\gamma_T)}{T} =: \frac{\Delta F}{T}.$$

Where $\Delta F$ is excess free energy. This shows that if the free energy of a state is small, that state is close to the thermal state.

To see why this is the case, we ask the following question: What is the probability that you measure a thermal state and get a state that looks like $\rho$? That's sort of like asking what the probability is that $\rho$ arises from fluctuation which, by Crooks fluctuation theorem, is given by $e^{-\Delta F/T} = 2^{D(\rho||\gamma_T)}$. So the relative entropy is saying something about how surprised you should be to see $\rho$ when you look at $\gamma_T$.

There is also another interpretation of $D(\rho||\gamma_T)$ in this case related to information removal and storage. Recall briefly Maxwell's Demon:

In this thought experiment, there is a box of gas particles with a partition in the middle, separating the left half from the right half. Further, there is a small hatch in the middle of this partition that can be open and shut by a demon in such as way as to not use any energy. If the demon opens that latch whenever a gas particle from the left side of the box is headed for it and closes it whenever a gas particle from the right is headed for it, eventually the gas particles will all end up on the right side of the box. This will have reduced the entropy and can therefore be used to perform work by opening the hatch and making the leftward motion of the gas particles do work. This however seems to violate the second law of thermodynamics.

The Landauer resolution to this paradox says that in fact this is not a violation because although the particles may be loosing entropy, the Demon is gaining information about which side of the box the gas is on and therefore is gaining entropy. In the act of gaining information, the Demon must also erase old information to make room for the new information. This erasure increased the entropy by at least as much as it is decreased by the collecting of the gas, giving us the minimal amount of work it costs to erase a bit, which by Landauer's erasure principle is $K_BT\ln(2)$. T

Along these lines, we $D(\rho||\gamma_T)$ as telling us how much space the state $\rho$ has to store information. The amount of work that can be extracted from state $\rho$ is given by $\Delta F = TD(\rho||\gamma_T)/\ln(2)$. Then, if we extract all the work we can from the state $\rho$ and use it to erase bits we can erase a total of

$$\frac{TD(\rho||\gamma_T)/\ln(2)}{K_BT\ln(2)} = \frac{D(\rho||\gamma_T)}{K_B}$$

bits. We can alternatively think of this operation as storing $D(\rho||\gamma_T)/K_B$ bits inside $\rho$.

**Second Law** We can also state a strong version of the second law of thermodynamics. For any channel $\mathcal{N}$ satisfying $\mathcal{N}(\gamma_T) = \gamma_T$ and any state $\rho$ we have

$$F(\mathcal{N}(\rho)) \leq F(\rho)$$

This is because any quantum channel can only decrease the relative entropy between $\rho$ and $\gamma_T$, so

$$F(\rho) - F(\gamma_T) = \frac{D(\rho||\gamma_T)}{\ln(2)} \geq \frac{D(\mathcal{N}(\rho)||\mathcal{N}(\gamma_T))}{\ln(2)} = F(\mathcal{N}(\rho)) - F(\gamma_T).$$

## 11.3 Application 3: Quantifying Entanglement

In this section we seek principled ways of quantifying entanglement between subsystems. We can first ask, what properties of this quantification migth make sense or be useful? To answer this questions, it will be useful to draw analogy to the case of trying to quantify someone's wealth. You can imagine that it might be easy to quantify the wealth of two people whose money is all in US dollars, we can simply count who has more. But what about comparing someone whose money is in US dollars to someone whose money is in Euros? Or someone whose wealth is in diamonds compared to someone whose wealth is in gold? It would be useful to have a single metric (such as a ”gold standard”) to compare these values on, such as converting them all to US dollars first. We also want this to be a fair comparison. If we suppose that in the process of converting from Euros to US dollars someone loses an excess amount of wealth or somehow gains extra wealth such that when they convert back to Euros, they end with significantly more or less money than they started with, then this hardly seems like a fair comparison. Therefore we want to be able to convert between currencies with a minimal ”exchange fee” so as to not significantly changing our wealth. Lastly, we would like it to be true that If we have our wealth in two different bank accounts, if we convert this wealth to US dollars, it doesn't matter if we convert it together or separately, we want to end up with the same amount of total US dollars at the end. This gives us the following properties:

1. Convertability
2. Small Conversion Fee
3. Additivity

These will be some of the properties that we may find useful when trying to judge a quantification scheme for entanglement.

### 11.3.1 Pure State Entanglement

We will begin by talking about quantifying the entanglement of pure states. Given a pure state $|\psi\rangle_{AB}$ the entanglement is quantified by the **entropy of entanglement**

$$S(A)_\psi = S(B)_\psi =: E$$

Explicitly, if $|\psi\rangle = \sum_i \sqrt{\lambda_i}\, |a_i\rangle \otimes |b_i\rangle$ then $E = H(\lambda)$.

Now we can ask Why is the entropy of entanglement a good measure of bipartite entanglement? We can imagine a conversion scheme, where different entangled states can be converted to the same ”currency” (in our case, this will be EPR pairs) through some set of operations. Further, we don't want to allow these operations to create new entanglement, just as we didn't want our conversions in the wealth example to create new wealth. Therefore, if we allow only local operators and classical communication

(LOCC), Bennett, Bernstein, Popescu, and Schumacher (arXiv 9511030) showed that we can transform our state $\psi$ as

$$\psi^{\otimes n} \to \Phi^{\otimes n(E-\delta)} \text{ (Entanglement Distillation)}$$
$$\Phi^{\otimes n(E+\delta)} \to \psi^{\otimes n} \text{ (Entanglement Dilution)}$$

where $\Phi = \frac{1}{\sqrt{2}}\left(|00\rangle + |11\rangle\right)$. This gives us our ”exchange rate” between any entangled state and the EPR state as $E = S(A) = S(B)$. Further, there is only a small exchange fee of $\delta$.Therefore, asymptotically up to LOCC we can think of $\psi$ as equalling $E$ copies of $\Phi$.

### 11.3.2 Mixed State Entanglement

What about the theory of bipartite entanglement for mixed states? We can try and do something analogous. Define the **distillable entanglement** $E_D(\rho)$ and **entanglement cost** $E_C(\rho)$ to be the max and min real numbers respectively such that

$$\rho^{\otimes n} \approx^{LOCC} \Phi^{E_D(\rho)} \text{ and} \tag{11.1}$$
$$\Phi^{E_C(\rho)} \approx^{LOCC} \rho^{\otimes n}. \tag{11.2}$$

Unfortunately, these definitions do not lead to the properties we discussed in the beginning. Some properties of these measures of entanglement include:

- They are not additive, so sometimes $E_D(\rho_1 \otimes \rho_2) > E_D(\rho_1) + E_D(\rho_2)$.
- There is no single letter formula known for these quantities. To get around this we can define the **entanglement of formation** $E_F(\rho)$ to be the minimum value of $\sum_i p_i S(\psi_i^A)$, taken over pairs of mixtures of pure states $(p_i, \psi_i)$ satisfying $\sum_i p_i \psi_i = \rho$. We have $E_C \leq E_F$, but sometimes this inequality is strict.
- Sometimes we have entangled states with $E_D = 0$ (”bound entanglement”), but we have no general theory of these states or why they occur.
- On the other hand, for any entangled state $E_C > 0$.
- $E_D \leq E_C$, and sometimes this inequality is strict, meaning we could lose ”entangledness” in converting back and forth between EPR pairs and certain states.

Let’s try and come up with a nicer measure of bipartite entanglement for mixed states.

### Relative Entropy of Entanglement

First define the **separable states** to be states in the set

$$\mathrm{Sep}(d_A, d_B) = \mathrm{conv}\left\{\alpha \otimes \beta : \alpha \in D_A, \beta \in D_B\right\}$$

where $D_A$ are $d_A \times d_A$ density matrices $D_B$ are $d_B \times d_B$ density matrices and

$$\mathrm{conv}(X) = \left\{\sum_{x \in X} p_x x : p_x > 0, \sum_x p_x = 1\right\}$$

is the convex hull of the points in $X$ (the smallest convex set that contains all of the points in $X$). These are our unentangled states. Unfortunately, it is NP hard to determine if a given state is separable.

Now, we can define the **relative entropy of entanglement**

$$E_R(\rho) = \min_{\sigma \in \mathrm{Sep}} D(\rho || \sigma)$$

This measure may be non-additive, so we also define the **regularized relative entropy of entanglement**

$$E_R^{\infty}(\rho) = \lim_{n \to \infty} \frac{1}{n} E_R(\rho^{\otimes n}) \leq E_R(\rho)$$

(and this inequality is sometimes strict).

Why is the relative entropy of entanglement nice? Define the **asymptotically non-entangling operations** (a family of operations that includes LOCC) to be channels $\Lambda_1, \Lambda_2, ... \Lambda_n : (A' \otimes B')^{\otimes n} \to (A \otimes B)^{\otimes n}$ with $\Lambda_n$ approximately sending separable states to separable states.

To make this definition precise define the **Rèyni divergences**

$$S_\alpha(A||B) = \frac{1}{\alpha - 1} \log(\mathrm{tr}[A^\alpha B^{1-\alpha}])$$

where as a few examples we have

$$\begin{aligned} S_1(A||B) &= S(A||B) \\ S_{1/2}(A||B) &= -2\log(F(A,B)) \\ S_\infty(A||B) &= \log \| B^{-1/2} A B^{-1/2} \|_\infty = \inf\{\lambda : A \leq 2^\lambda B\} \end{aligned}$$

Now channels $\Lambda_1, \Lambda_2, ... \Lambda_n$ are asymptotically non-entangling if

$$\forall \rho, \sigma \in \mathrm{Sep} : S_\infty(\Lambda_n(\rho^{\otimes n}) || \sigma) \leq \epsilon_n$$

with $\epsilon_n \to 0$ as $n \to \infty$. This gives us our precise definition of these operations that we will now use to examine the regularized relative entropy of entanglement.

**Theorem 7 (Brandao-Plenio arXiv:0710.5827)** *Up to asymptotically non-entangling operations*

$$\rho^{\otimes n} \leftrightarrow \Phi^{\otimes n E_R^\infty(\rho)}$$

This is exactly the type of conversion we were looking for in our metric of entanglement.

We note that a similar result holds in thermodynamics. Define **thermal operations** to be channels

$$\mathcal{N}(\rho) = \mathrm{tr}_E \left[ V \left( \rho_S \otimes \frac{\exp(-\beta H_E)}{\mathrm{tr}[\exp(-\beta H_E)]} \right) V^\dagger \right]$$

with $[V, H_S \otimes I + I \otimes H_E] = 0$. $H_S$ is the system Hamiltonian while $H_E$ is the environment (or bath) Hamiltonian. These are the operations that are free if the thermal states are free. In other words, they do not create any free energy. Let $\gamma_T = e^{-\beta H_S}/\mathrm{tr}[e^{-\beta H_S}]$ be the thermal state of the system. Under thermal operations, we can transform a state $\rho$ into a state $\sigma$ at a rate $D(\rho||\gamma_T)/D(\sigma||\gamma_T)$. In this way, we can think of entanglement and non-thermal states as resources.

We will now conclude by sketching the proof of Brandao-Plenio.

First, we compute $E_R(\Phi^{\otimes n})$. We do this via Stein's Lemma and ote that the optimal measurement to distinguish any state from the EPR state is $M = \Phi^{\otimes n}$. Taking $\sigma \in \mathrm{Sep}$ we have

$$\begin{aligned}
\max_\sigma \mathrm{tr}[M\sigma] &= \max_{|\alpha\rangle, |\beta\rangle} \left| \langle \Phi |^{\otimes n} |\alpha\rangle |\beta\rangle \right|^2 \\
&= \max_{|\alpha\rangle, |\beta\rangle} |\langle \alpha | |\beta\rangle|^2 / 2^n = 2^{-n}
\end{aligned}$$

which gives $E_R(\Phi^{\otimes n}) = n$. This is obviously the dsired result, since it tells us that $n$ EPR states are worth $n$ EPR states worth of entanglement.

Now, for any state $\rho$, $S_\infty(\rho||\mathrm{Sep}) = \lambda$ implies that there exists a $\sigma \in \mathrm{Sep}$ with $\rho^{\otimes n} \leq 2^\lambda \sigma$. Equivalently, $2^{-\lambda}\rho^{\otimes n} \leq \sigma$. Then we can write

$$\sigma = 2^{-\lambda}\rho^{\otimes n} + (I - 2^{-\lambda})\gamma$$

for some density matrix $\gamma$. Define an asymptotically non-entangling operation where $\Lambda_n$ is the either the measurement $\Phi^{\otimes nR}$ and outputs $\rho^{\otimes n}$ or the measurement $I - \Phi^{\otimes nR}$ and outputs $\gamma$. The outcome of this measurement on $\Phi^{\otimes nR}$ is $\phi^{\otimes n}$ and the outcome of the measurement on any separable state is $2^{-\lambda}\rho^{\otimes n} + (I - 2^{-\lambda})\gamma = \sigma$. So the measurement is asymptotically non-entangling and maps $\Phi^{\otimes nR}$ to $\rho^{\otimes n}$.

To map in the other direction we use the optimal test distinguishing $\rho^{\otimes n}$ from Sep.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 12: October 8, 2020

*Lecturer: Aram Harrow* *Scribe: Yuan Lee, Adam Bene Watts*

Previously, we saw that feedback (and hence shared randomness) has no effect on the classical channel capacity. However, quantum mechanics *does* affect the way we think about channel coding, and it shows up in a variety of ways. Some examples follow.

1. **Entanglement assistance:** using entanglement as a resource to transmit classical messages. This can increase the capacity of a quantum or classical channel (i.e. superdense coding, depolarizing channel example on pset 5).
2. **Entangled inputs:** even if Alice and Bob don't share entanglement, Alice can entangle her inputs across many uses of the channel. This extends the idea of correlations in classical codebooks by using entanglement resources.
3. **Quantum capacities:** where the goal is to transmit quantum messages.

Quantum capacities are further related to secret key capacities through quantum key distribution (QKD): transmitting one qubit allows the sender and recipient to share one secret bit.

## 12.1 Resource Notation

We keep track of communication resources using the following scheme.

- $[c \to c]$ or $[c \leftarrow c]$ = noiseless transmission of one cbit (classical bit).
- $[cc]$ = one rbit (shared random bit).
- $[q \to q]$ or $[q \leftarrow q]$ = noiseless transmission of one qubit.
- $[qq]$ = one bit of shared entanglement ("ebit"): specifically, the Bell pair $|\Phi\rangle = (|00\rangle + |11\rangle)/\sqrt{2})$.
- $\langle N \rangle$ = one channel use $N_{A' \to B}$.
- $\langle \rho \rangle$ = one copy of $\rho$.

We say that $a \geq b$ if the combination of resources in $a$ can generate the resources in $b$ using a protocol with asymptotically vanishing error and inefficiency (note that we only care about the rate achieved in the limit of many uses of $a$). This definition will lets us define a partial order on resources.

The achievability portion of the channel capacities we saw before can now be written using this resource notation.

- classical capacity: $\langle N \rangle \geq C(N)[c \to c]$.
- quantum capacity: $\langle N \rangle \geq Q(N)[q \to q]$.
- entanglement-assisted capacities:

$$\langle N \rangle + \infty[qq] \geq C_E(N)[c \to c] \text{ and } \langle N \rangle + \infty[qq] \geq Q_E(N)[q \to q].$$

- distillable entanglements:

$$\langle \rho \rangle + \infty[c \to c] \geq E_{D,1}(\rho)[qq] \text{ (1-way distillable entanglement) and}$$
$$\langle \rho \rangle + \infty[c \to c] + \infty[c \leftarrow c] \geq E_{D,2}(\rho)[qq] \text{ (2-way distillable entanglement).}$$

As an example of the partial ordering on resources, we give the following diagram of relationships, where $a \to b$ means $a \geq b$ and entries without arrows between them are incomparable.

$$\begin{tikzcd}
& {[q \to q]} \arrow[dl] \arrow[dr] & & {[q \leftarrow q]} \arrow[dl] \arrow[dr] & \\
{[c \to c]} \arrow[drr] & & {[qq]} \arrow[d] & & {[c \leftarrow c]} \arrow[dll] \\
& & {[cc]} & &
\end{tikzcd}$$

We can also write describe quantum teleportation and superdense coding using resource notation.

- **teleportation** can be written as $[qq] + 2[c \to c] \geq [q \to q]$.
- **superdense coding** can be written as $[q \to q] + [qq] \geq 2[c \to c]$.

(Resource notation obscures the fact that these protocols work non-asymptotically.)

## 12.2 Assorted Topics

### 12.2.1 Channel Simulation

$\langle N\rangle \geq C(N)[c \to c]$ implies that we can simulate $\approx nC(N)$ copies of $[c \to c]$ using $n$ uses of channel $N$. Could we simulate $N$ using $[c \to c]$ instead? The classical reverse Shannon theorem says we can (note $N$ is classical channel):

$$C(N)[c- > c] + \infty[cc] \geq \langle N\rangle.$$

But what does it mean precisely simulate a channel?

We say that we can simulate the target channel $N$ if our protocol produces a channel $M$ that is close to $N$. Two metrics for channels are:

- **diamond norm** $\|M - N\|_\diamond = \|\mathrm{id} \otimes M - \mathrm{id} \otimes N\|_{1\to 1}$;
- $1 \to 1$ **norm** $\|M - N\|_{1\to 1}$, defined by

  $$\|\mathcal{E}\|_{1\to 1} = \sup_X \frac{\|\mathcal{E}(X)\|_1}{\|X\|_1}.$$

  If $\mathcal{E}$ is the difference between two channels, an equivalent definition is $\|\mathcal{E}\|_{1\to 1} = \max_\rho \|\mathcal{E}(\rho)\|_1$.

The diamond norm gives the strongest condition on “closeness”, whereas the 1 → 1 norm gives a weaker condition. Intuitively, this is because some channels can be better distinguished by feeding in states entangled with some reference system.

We can use either of these metrics to define the accuracy of our channel simulation. The diamond norm gives the strongest constraint, and bounds the ammount of error we can incur by replacing a channel $\mathcal{N}$ by a simulation $\mathcal{M}$ in some protocol. Both the diamond norm and 1 to 1 norm apply to blind inputs, where Alice does not have a classical description of her state. We can also consider the case where Alice has a description of her state $\rho$, and Bob wants to construct the state $\mathcal{N}(\rho)$. This gives an even weaker condition on “closeness”.

### 12.2.2 Resource Arithmetic

For any positive $\epsilon$, we cannot do

$$(2 - \epsilon)[c \to c] + 1000[qq] \geq [q \to q]. \tag{12.1}$$

(The coefficient of 1000 is illustrative; take it to be an arbitrary large number.)

**Proof.** Equation (12.1) violates the no-signalling theorem through superdense coding. If a protocol achieves (12.1), then we can use one extra ebit to obtain

$$(2-\epsilon)[c \to c] + 1001[qq] \geq [q \to q] + [qq] \geq 2[c \to c].$$

If this were possible, then for sufficiently large $n$, there will be two possible received messages corresponding to the same input message, distinguished only by the ebit $[qq]$. If we fix this input message, then we can use entanglement to transmit a classical bit, contradicting the no-signalling theorem.

### 12.2.3 Remote State Preparation

Let "$\psi$" refer to the classical description of the $n$-qubit state $\psi$. Then, *remote state preparation* refers to a protocol in which Bob prepares the state $\psi$ with Alice's help. Alice has access to the classical description of $\psi$, $n$ ebits (shared with Bob) and $n(1+\delta)$ cbits (for some $\delta > 0$):

$$\text{``}\psi\text{''} + n[qq] + n(1+\delta)[c \to c] \Rightarrow \psi.$$

The proof of remote state preparation is relatively involved.

Remote state preparation can be thought of as a simulation of the identity channel with visible inputs. Comparing the resource cost of this to the resource cost of simulating the identity channel on blind inputs (i.e. generating the resource $[q \to q]$) derived above shows the difference in resource cost between visible and blind simulation.

## 12.3 Entangled Inputs

The HSW theorem gives the classical capacity of a CQ channel.

For a general quantum channel $N$, the capacity with product state inputs only is

$$\chi(N) = \max_{\rho} I(X;B)_\rho$$

where $\rho$ is chosen from the set of states with form $\rho = \sum_x p(x) \left|x\right\rangle \left\langle x\right|_X \otimes N_{A' \to B}(\psi^x_{A'})$. This bound essentially comes from the original HSW theorem and the observation that, even when communicating through a quantum to quantum channel $N$, Alice must make a classical choice of a message $x$ to transmit to Bob, then send an associated pure state $\psi^x_{A'}$ through the channel to Bob. (She could also send mixed states but we could always decompose those into pure states and add extra labels, which would only increase channel capacity).

However, the inputs to $N^{\otimes n}$ over $n$ channel uses can be entangled. The classical channel capacity of $N$ is defined to be the regularization of $\chi$:

$$C(N) = \lim_{n\to\infty} \frac{1}{n}\chi(N^{\otimes n}).$$

This capacity $C$ is generally hard to compute.

**Facts about $\chi$ and $C$.**

1. Computing $\chi$ is an NP-complete optimization problem (scaling in terms of the dimension of the channel.)
2. $C \geq \chi$ (since we can just use product states). Sometimes, $C > \chi$.
3. The complexity of $C$ is unknown. It could be polynomial, it could be uncomputable.
4. $\chi$ is easy to compute for CQ channels. $C_E$ is also easy to compute.
5. Let an additivity violation for a capacity $\chi$ refer to the existence of channels $N_1$ and $N_2$ such that $\chi(N_1 \otimes N_2) > \chi(N_1) + \chi(N_2)$. Then Shor (quant-ph/0305035) showed that

   $$\chi \text{ additivity violation} \Leftrightarrow E_F \text{ additivity violation} \Leftrightarrow S_{\min} \text{ additivity violation},$$

   where $S_{\min}(N) = \min_\rho S(N(\rho))$ is the entropy of the least-mixed output. Hastings (0809.3972) later proved additivity violation.
6. $\chi$ is known to be additive in the following cases.
   - **Entanglement-breaking channels**. We say that $N$ is entanglement-breaking if $(\mathrm{id}\otimes N)(\rho) \in \mathrm{Sep}$ for all inputs $\rho$. Equivalently, $N$ is entanglement-breaking iff it can be written as a measure-and-prepare (or QCQ) channel.
   - **Depolarizing channels** $N(\rho) = (1-p)\rho + pI/d$ for some probability $p$.
   - **Erasure channels** $N(\rho) = (1-p)\rho + p\,|e\rangle\langle e|$, where $|e\rangle$ is an erasure flag.
   - **Unital qubit channels** $N(I/d) = I/d$.
   - **Purely lossy bosonic channels**, where the output mode $a_k'$ can be described in terms of input modes $a_k$ and $b_k$ as $a_k' = \sqrt{\eta_k}a_k + \sqrt{1-\eta_k}b_k$.

- **Hadamard channels**. If the Stinespring representation of $N$ is $N(\rho) = \mathrm{tr}_E V_{A' \to BE} \rho V^\dagger_{A' \to BE}$, then the complement of $N$ is $N^c(\rho) = \mathrm{tr}_B V_{A' \to BE} \rho_{A'} V^\dagger_{A' \to BE}$. Then $N$ is a Hadamard channel iff $N^c$ is an entanglement-breaking channel.

We conclude with a brief outline of the main ideas of Hastings' proof of superadditivity. Hastings showed existence of a channel $\mathcal{N}$ which satisfied

$$S_{min}(\mathcal{N} + \overline{\mathcal{N}}) \leq S_{min}(\mathcal{N}) + S_{min}(\overline{\mathcal{N}})$$

(the channel $\overline{\mathcal{N}}$ is obtained by taking the complex conjugate of everything in $\mathcal{N}$).

The channel $\mathcal{N}$ is defined act randomly on the state $\rho$ with one of $D$ possible unitaries ($D$ is a constant independent of the dimension of the state): $\mathcal{N}(\rho) = \sum_{i=1}^D U_i \rho U_i^\dagger$.

To understand why entanglement state inputs to the channel $\mathcal{N} + \overline{\mathcal{N}}$ can lead to a lower output entropy, consider sending the maximally mixed state $\Phi$ through the channel $\mathcal{N} \otimes \overline{\mathcal{N}}$. Then

$$\begin{aligned} \mathcal{N} \otimes \overline{\mathcal{N}}(\Phi) &= \frac{1}{d^2} \left( \sum_i (U_i \otimes \overline{U_i}) \Phi (U_i \otimes \overline{U_i}) + \sum_{i \neq j} (U_i \otimes \overline{U_j}) \Phi (U_i \otimes \overline{U_j}) \right) \\ &= \frac{1}{d} \Phi + \frac{1}{d^2} \left( \sum_{i \neq j} (U_i \otimes \overline{U_j}) \Phi (U_i \otimes \overline{U_j}) \right), \end{aligned}$$

where we used that $(I \otimes \overline{U_i})\Phi = (\overline{U_i}^\top \otimes I)\Phi = \left(U_i^{-1} \otimes I\right)\Phi$. From this, a not-to-hard calculation shows

$$S_{min}(\mathcal{N} \otimes \overline{\mathcal{N}}) \leq S(\mathcal{N} \otimes \overline{\mathcal{N}}(\Phi)) \leq 2\ln(D) - \frac{\ln(D)}{2}.$$

A very hard calculation shows that

$$S_{min}(\mathcal{N}) \geq \ln(D) - \frac{C}{D} - D^{O(1)} \sqrt{\frac{\ln(N)}{N}}$$

(C is a constant, $N$ is the dimension of the channel), which proves the result.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 13: October 15, 2020

*Lecturer: Aram Harrow* *Scribe: Thiago Bergamaschi, Yuan Lee*

We begin by exploring the intuition for the connections between the Holevo $\chi$ quantity, the minimum entropy and the entanglement of formation in Shor's 2003 paper quant-ph/0305035, following the discussion in the previous lecture.

## 13.1 Sub-additivity of $S_{min}$ $\Rightarrow$ Super-additivity of $\chi$

Given an ensemble of states $\{p_i, \rho_i\}$ with average state $\bar{\rho} = \sum p_i \rho_i$, the Holevo $\chi$ quantity is

$$\chi(N) = S(N(\bar{\rho})) - \sum p_i S(N(\rho_i)) \tag{13.1}$$

by definition.

We can bound $\chi$ from above using

$$\chi(N) \leq S(N(\bar{\rho})) - S_{min}(N) \leq S_{max}(N) - S_{min}(N) \leq \log d_B - S_{min}(N) \tag{13.2}$$

Shor showed that for every channel $N$, one can construct a channel $N'$ that makes the inequalities above tight. In particular, if $N$ has dimension $d_B$,

$$\chi(N') = \log d_B - S_{min}(N) \tag{13.3}$$

The construction is quite straightforward. After the application of the quantum channel $N$, apply a classically-controlled random Pauli operator $\sigma_x$, so that $N'(\rho) = \sigma_x N(\rho) \sigma_x^\dagger$. In this manner, the first term in equation (13.1) is $S(N(\bar{\rho})) = \log d_B$, because

$$N' \left( \sum_x \frac{1}{d_B^2} |x\rangle\langle x| \otimes \rho \right) = \sum_x \frac{1}{d_B^2} \sigma_x N(\rho) \sigma_x^\dagger = \frac{\mathbb{I}}{d_B} \tag{13.4}$$

Moreover, the second term is $S_{min}(N') = S_{min}(N)$. It follows that if $S_{min}$ is subadditive, then $\chi$ is superadditive.

The other direction is non-trivial.

## 13.2 Renyi Entropies

$S_{min}$ is generally computationally more tractable than $\chi$, as we can view them as the limit of Renyi entropies. Recall

$$S_\alpha(\rho) = \frac{1}{1-\alpha}\text{Tr}[\rho^\alpha] \tag{13.1}$$

There are a few particular cases of $\alpha$ to highlight: $S_0 = \log \text{rank}\, \rho, S_\infty = -\log ||\rho||_\infty$ and $S_1(\rho) = S(\rho)$, the standard Von Neumann entropy. Analogously, we can define the min Renyi Entropy via

$$S_{\alpha,min}(N) = \min_\psi S_\alpha(N(\psi)) = \frac{\alpha}{1-\alpha}\log ||N||_{1\to\alpha} \tag{13.2}$$

where $||N||_{\beta\to\alpha}$ is the “beta to alpha norm” defined as follows

$$||N||_{\beta\to\alpha} = \sup\frac{||N(X)||_\alpha}{||X||_\beta} \tag{13.3}$$

Finding $S_{\alpha,min}$ is still a hard optimization problem, but $S_{\alpha,min}$ is more helpful to us because the norms it is related to obey useful inequalities.

## 13.3 The Connection to the Entanglement of Formation

We can analogously extend the Holevo information of a state $\rho$, by decomposing over an ensemble of pure states $\{p, \phi\}$ that averages $\rho$

$$\chi(N,\rho) = \max_{\{p,\phi\} \text{ s.t. } \sum_x p_x\phi_x = \rho} S(N(\rho)) - \sum p_x S(N(\phi_x)) \tag{13.1}$$

where to conclude $\chi(N) = \max_\rho \chi(N,\rho)$. If we consider applying the Stinespring dilation theorem to $N$ s.t. $N(\omega) = \text{Tr}_E(V\omega V^\dagger)$, then $S(N(\phi_x))$ is simply the entanglement of $V|\phi_x\rangle$, and it follows

$$\chi(N,\rho) = S(N(\rho)) - E_F(V\rho V^\dagger) \tag{13.2}$$

We might be concerned that not all entanglements of formation $E_F(V\rho V^\dagger)$ correspond to the *minimum* average entropy of the corresponding channel $N$. However, the MSW correspondence states that

$$E_F\left(\rho^{BE}\right) = \min_{\{p,\phi\} \text{ s.t. } \sum_x p_x\phi_x = \rho} \sum_x p_x S\left(\text{tr}_E \phi_x^{BE}\right)$$

for any bipartite state $\rho$.

Applying it to the Stinespring dilation of $N_{A\to B}$, we have

$$E_F(V\rho V^\dagger) = \min_{\{p,\phi\} \text{ s.t. } \sum_x p_x\phi_x=\rho} \sum_x p_x S\left(\mathrm{tr}_E V\phi_x V^\dagger\right)$$

which guarantees that equation (13.2) holds.

## 13.4 Entanglement-Assisted Capacity

The discussion of $S_{min}$ and $\chi$ above only delays the pain of performing the optimization problem over $\{p,\phi\}$ such that $\sum_x p_x\phi_x = \rho$. Now we turn to the more well-understood problem of entanglement-assisted capacities.

We want to analyze the additivity of the entanglement-assisted capacity of two independent channels $C_E(N_1 \otimes N_2)$. Define systems $A_1', A_2'$ upon which $N_1, N_2$ act, respectively, and consider an environment system $A$

$$C_E(N_1 \otimes N_2) = \max I(A : B_1B_2)_\tau, \tau = (\mathbb{I}_A \otimes N_1^{A_1'\to B_1} \otimes N_2^{A_2'\to B_2})(\phi^{AA_1'A_2'}) \tag{13.1}$$

$$= \max I(A : B_1B_2)_\psi, |\psi\rangle = (\mathbb{I}_A \otimes V_1^{A_1'\to B_1E_1} \otimes V_2^{A_2'\to B_2E_2})|\phi\rangle \tag{13.2}$$

note the distinction between the two definitions, where in the second we purify the two systems independently s.t. they are separable in $\psi$ but not in $\tau$. We will use this independence later. It follows now that we can apply the chain rule sequentially

$$I(A : B_1B_2)_\psi = I(A : B_1) + I(A : B_2|B_1) = \tag{13.3}$$

$$= I(A : B_1) + I(AB_1 : B_2) - I(B_1 : B_2) \leq I(A : B_1) + I(AB_1 : B_2) \tag{13.4}$$

as the mutual information is non-negative. Let us consider the terms above independently, starting by $I(AB_1 : B_2)$. Intuitively, if it was helpful to include $B_1$ in addition to $A$, we could have included it into the definition of the 'environment system' WLOG. In this manner, $I(AB_1 : B_2) \leq I(AB_1E_1 : B_2)$, and symmetrically for $1 \leftrightarrow 2$. We conclude

$$C_E(N_1 \otimes N_2) \leq I(AB_1E_1 : B_2) + I(AB_2E_2 : B_1) \leq C_E(N_1) + C_E(N_2) \tag{13.5}$$

Finally, note that this upper bound is always achievable as we can run the channels independently. We conclude

$$C_E(N_1 \otimes N_2) = C_E(N_1) + C_E(N_2) \tag{13.6}$$

and therefore we conclude $C_E$ is additive and has a single-letter formula that is concave in $\rho$. Moreover, through superdense coding and quantum teleportation (problem set 5, problem 1) it determines the quantum entanglement-assisted capacity $Q_E = C_E/2$ as well.

We can contrast these nice properties of the entanglement-assisted capacities ($C_E$ additive and $C_E = 2Q_E$) with the difficulty of the unassisted capacities ($C$, $Q$). We only know that $Q \leq C$, but this bound can have large gaps. For instance, the completely dephasing channel has $C = 1$ but $Q = 0$. On the other hand, the noiseless channel has $Q = C$. This makes it difficult to think of channels as equivalent resources in the absence of free entanglement.

## 13.5 Quantum Reverse Shannon Theorem and Embezzling States

The additivity of $C_E$ and the reversibility of the quantum Shannon theorem allows us to think of channels as equivalent resources, associated with a resource theory. In particular, reversibility (defined formally below) allows us to convert between channels at a common "exchange rate".

The quantum reverse Shannon theorem states that any quantum channel can be simulated by an 'unlimited amount' of shared entanglement and $C_E$ classical bits, where $C_E$ is the entanglement-assisted classical capacity of the channel. In informal resource notation,

$$\text{unlimited entanglement} + C_E[c \to c] \geq \langle N \rangle \tag{13.1}$$

The lecturer traces a key distinction here between 'unlimited entanglement' and $\infty[qq]$, an arbitrary amount of EPR pairs. The key intuition is that the channel simulation may consume a different amount of EPR pairs for different inputs, and therefore it doesn't suffice to feed some amount of EPR pairs to the protocol. Instead, *embezzling states* are bipartite states that allow the removal of a small amount of entanglement under local operations into an additional set of registers, while the original state remains approximately the same. That is, heuristically,

$$|\Gamma\rangle_{AB} \to\approx |\Gamma\rangle_{AB} \otimes |\psi\rangle_{A'B'} \tag{13.2}$$

where the A'B' registers are much smaller than AB. A motivating example is the following state.

$$|\Gamma\rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |\Phi_2\rangle^{\otimes i} \otimes |00\rangle^{\otimes n-i} |ii\rangle \tag{13.3}$$

Note that if we define $\Gamma'$ based on the removal of the first Bell pair $\Phi_2$, i.e.

$$|\Gamma'\rangle = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} |\Phi_2\rangle^{\otimes i} \otimes |00\rangle^{\otimes n-i} |ii\rangle \tag{13.4}$$

then the fidelity $F(\Gamma, \Gamma') = 1 - \frac{1}{n}$ and we have 'stolen an EPR pair'.

Another example of an embezzling state is

$$|\Psi\rangle \propto \sum_{i=1}^{N} \frac{1}{\sqrt{i}} |ii\rangle$$

for some finite $N$. Embezzling entanglement then looks like

$$(U^A \otimes V^B) |\Psi\rangle^{AB} |00\rangle^{AB} \approx |\Psi\rangle^{AB} |\Phi_2\rangle^{AB}$$

for some local unitaries $U^A$ and $V^B$.

We can show that there exist local unitaries $U, V$ such that $F((U \otimes V) |\Psi\rangle |00\rangle, |\Psi\rangle |\Phi_2\rangle) \geq 1 - 1/\log n$. Let $\sum_{i=1}^{N} 1/i = C_N$. The Schmidt coefficients of $|\Psi\rangle |00\rangle$ are

$$\frac{1}{\sqrt{C_N}}, \frac{1}{\sqrt{2C_N}}, \frac{1}{\sqrt{3C_N}}, \frac{1}{\sqrt{4C_N}}, \dots \frac{1}{\sqrt{NC_N}}, 0, \dots 0$$

whereas the Schmidt coefficients of $|\Psi\rangle |\Phi_2\rangle$ are

$$\frac{1}{\sqrt{2C_N}}, \frac{1}{\sqrt{2C_N}}, \frac{1}{\sqrt{4C_N}}, \frac{1}{\sqrt{4C_N}}, \frac{1}{\sqrt{6C_N}}, \frac{1}{\sqrt{6C_N}}, \dots, , \frac{1}{\sqrt{2NC_N}}, \frac{1}{\sqrt{2NC_N}}$$

Therefore, the maximum fidelity is

$$\begin{aligned}
&\left(\frac{1}{\sqrt{C_N}}, \frac{1}{\sqrt{2C_N}}, \dots \frac{1}{\sqrt{NC_N}}, 0, \dots 0\right) \cdot \left(\frac{1}{\sqrt{2C_N}}, \frac{1}{\sqrt{2C_N}}, \frac{1}{\sqrt{4C_N}}, \frac{1}{\sqrt{4C_N}}, \dots \frac{1}{\sqrt{2NC_N}}\right) \\
&\geq \frac{1}{C_N}\left(\frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{N}\right) \\
&\geq \frac{1}{C_N}\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{\lfloor N/2 \rfloor}\right) \\
&\geq \frac{\ln N/2}{\ln N} = 1 - \frac{1}{\log N}
\end{aligned}$$

Note that entanglement embezzlement preserves the original superposition across the bipartite state $|\Psi\rangle$, which is crucial for the quantum reverse Shannon theorem.

## 13.6 Quantum Capacity

In resource notation, we define the quantum capacity by

$$\langle N\rangle \geq Q(N)[q \to q] \tag{13.1}$$

In general, $Q \leq Q_F \leq Q_2$, that is feedback and two-way channels increase the capacity of quantum information over the channel, however, sending additional classical communication $[c \to c]$ does not help. Mathematically, the quantum capacity is defined by the maximum amount distillable entanglement that can be generated with the channel

$$Q(N) = \max_{\psi_{AA'}} E_D((\mathbb{I}_A \otimes N_{A'\to B})\psi_{AA'}) \tag{13.2}$$

The Choi-Jamiolkowski state $\omega(N)$ is

$$\omega(N) = (\mathbb{I}_A \otimes N_{A'\to B})(\Phi_{AA'}) = \frac{1}{d_A}\sum_{ij} |i\rangle\langle j| \otimes N(|i\rangle\langle j|) \tag{13.3}$$

where $\Phi_{AA'} = \frac{1}{\sqrt{d_A}}\sum_i |ii\rangle$ is the maximally mixed state. This state presents an interesting interpretation of the channel, as the mapping $N \to \omega(N)$ is an isomorphism (known as the *Choi-Jamiolkowski isomorphism*). We can show that this mapping is isomorphic by identifying the inverse map: conditioning on the first subsystem of $\omega(N)$, we obtain $N(|i\rangle\langle j|)$ for every basis element $|i\rangle\langle j|$, which suffices to define the channel $N$.

We can use $\omega$ to simulate $N$ as follows. Consider three registers $E, A, A'$, where $E$ holds a state $\rho$ and $A, A'$ share the maximally mixed state $\Phi_{AA'}$. Consider the quantum circuit defined by feeding $A'$ through the quantum channel $N$, and a Bell state measurement is jointly made on the registers $E, A$. If the bell state measurement returns a string $j$, then the state resulting on the register $A' \to B$ is $N(\sigma_j \rho \sigma_j^\dagger)$. In this manner, $j = 0$ with probability $d_A^{-2}$, and then $N(\sigma_j\rho\sigma_j^\dagger) = N(\rho)$. It follows $\omega(N)$ can simulate $N$ with probability $d_A^{-2}$ and in this manner,

$$E_D(\omega(N)) > 0 \iff Q(N) > 0 \tag{13.4}$$

Unfortunately, it is still largely unknown when $Q(N) = 0$. A case that $Q(N) = 0$ is when $N$ is entanglement-breaking, or equivalently when $\omega(N) \in \text{Sep}$ is separable.

We can generalize entanglement-breaking channels in two different ways. One way to generalize the entanglement-breaking property is to consider antidegradable channels.

We say $N$ is *antidegradable* if there exists some map $\varepsilon$ such that $N = \varepsilon \circ N^c$ (i.e. Bob gets less information than Eve). Conversely, we say that $N$ is *degradable* if there exists some map $\varepsilon$ such that $N^c = \varepsilon \circ N$ (i.e. Bob gets more information than Eve). For example, the erasure channel with erasure probability $p$ is degradable for $p \leq 1/2$ and antidegradable for $p \geq 1/2$. In general, however, not all channels are either degradable or antidegradable.

We can show using the no-cloning theorem that antidegradable channels also have zero quantum capacity (without classical feedback). Interestingly, degradable channels have additive capacity.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 14: October 20, 2020

*Lecturer: Aram Harrow* *Scribe: Leon Ding, Thiago Bergamaschi*

# 14.1 Quantum Capacity formula

For a given quantum channel $N_{A\to B}$, consider the definition of an environment system under the Stinespring representation $N(\rho) = \mathrm{Tr}_E[V\rho V^\dagger]$. We define the **coherent information** as

$$I_C(\rho, N) = S(N(\rho)) - S(N^c(\rho)) = S(B) - S(E) \tag{14.1}$$

Where the superscript $c$ denotes tracing out the complement subspace. For example, if in $N(\rho)$ we trace out subspace $E$, then in $N^c(\rho)$ we trace out subspace $B$. Under this definition, we can now define the **quantum capacity** according to the LSD theorem (Lloyd, Shor, Devetak):

$$Q(N) = \lim_{n\to\infty} \frac{1}{n} \max_\rho I_c(\rho, N^{\otimes n}) \tag{14.2}$$

$$= \max_\rho I_c(\rho, N) \qquad \text{if } N \text{ is degradable} \tag{14.3}$$

Let us consider the definition of $I_c$ through the purification of the initial state. In particular, let $\phi_{AA'}$ be a pure state, and consider the subsystems $B$, $E$ resulting of feeding $A'$ through the channel $N$. Under the tri-partite state $\tau_{ABE}$,

$$I_c = S(B)_\tau - S(E)_\tau = S(B) - S(AB) = -S(A|B) = \frac{I(A:B) - I(A:E)}{2} \tag{14.4}$$

Likewise we can consider the entanglement of distillation $E_D(\rho)$

$$E_D(\rho_{AB}) = \lim_{n\to\infty} \frac{1}{n} \max_{\Lambda: A_1\cdots A_n \to AE'} [S(B_1\cdots B_n) - S(AB_1\cdots B_n)] \tag{14.5}$$

We can define metric of capacity using an arbitrary penalty on the conditional entropy between $A$ and $B$ called the **Hare-brained capacity**:

$$C_{\mathrm{HB}}(N) = \max H(B) - 10H(B|A) \tag{14.6}$$

While this penality is arbitrary, it still satisfies

$$\lim_{n\to\infty} \frac{1}{n} \max_\rho C_{\mathrm{HB}}(N^{\otimes n}, \rho) = C(N) \tag{14.7}$$

## 14.2 Understanding the Capacity Formula

1. $N$ is antidegradable. If $E = E_B E_E$, $I(A:E) = I(A:E_B) + I(A:E_E|E_B) \geq I(A:E_B) = I(A:B)$

   $I_C \leq 0$. Though this statement does not hold if allowing feedback.

2. Consider a random Pauli channel, which applies a random Pauli matrix with some respective probability. $N(\rho) = (1 - p_x - p_y - p_z)\rho + p_x X\rho X + p_y Y\rho Y + p_z Z\rho Z$

   Applying this channel,

   $$(I \otimes N)\Phi = p_I \Psi_0 + p_x \Psi_1 + p_y \Psi_2 + p_z \Psi_3 \tag{14.1}$$

   where $|\Psi_i\rangle = (I \otimes \sigma_i)|\Psi\rangle$

   A purification of this density matrix is the wavefunction

   $$\sqrt{p_I}\,|\Psi_0\rangle_{AB}|0\rangle_E + \sqrt{p_X}\,|\Psi_1\rangle_{AB}|1\rangle_E + \sqrt{p_Y}\,|\Psi_2\rangle_{AB}|2\rangle_E + \sqrt{p_Z}\,|\Psi_3\rangle_{AB}|3\rangle_E \tag{14.2}$$

   which has $S(B) = 1$ and $S(E) = H(\vec{p})$.

   A special case of this is the **depolarizing channel** $D_p$, with $S(E) = H_2(p) + p\log 3$.

   "Hashing bound" cf. hashing method to check if $x \overset{?}{=} y$. We choose some random function $f \to \{0,1\}^k$ and check if $f(x) \overset{?}{=} f(y)$. If $x = y$, we necessarily have $f(x) = f(y)$, and if $x \neq y$, the probability that $f(x) = f(y)$ can be be shown to be small $\Pr[f(x) = f(y)] \sim 2^{-k}$

3. Sometimes preprocessing helps

   $$\rho = \Phi_{A,B} \otimes \left(\frac{I}{2}\right) A_2 \tag{14.3}$$

   $I_C = 0$, preprocessing results in $\to 1$ or $I_C(I/2, D_p) = 1 - H_2(p) - p\log 3$.

4. Sometimes entangled inputs help. `quant-ph/9706061` for $p \approx 0.19$, $I_C(I/2, D_p) < \frac{1}{5} I_C(\frac{|00000\rangle\langle 00000| + |11111\rangle\langle 11111|}{2}, D_p^{\otimes 5})$

5. **Superactivation** $\exists N_1, N_2$ s.t. $Q(N_1) = Q(N_2) = 0$, but $Q(N_1 \otimes N_2) > 0$

   An example of this is with $N_1 = 50\%$ erasure channel and $N_2$ a PPT channel with private $C \propto \rho > 0$(???). This satisfied $Q(N_1 \otimes N_2) > 0$ according to `Smith-Yard, Science 2008, arxiv:0807.4935`.

## 14.3 PPT Channels

The **partial transpose** of a density matrix is defined as

$$\rho^\Gamma = (I \otimes T)\rho \tag{14.1}$$

Where for example

$$(|i\rangle\langle j| \otimes |k\rangle\langle l|)^\Gamma = |i\rangle\langle j| \otimes |l\rangle|k\rangle \tag{14.2}$$

Let us start to build some intuition on this operation. Let PPT be the set of density matrices that under partial transpose remain positive semi-definite, i.e.

$$\text{PPT} = \{\rho : \rho^\Gamma \geq 0\} \tag{14.3}$$

Since all psd matrices are symmetric, $\rho^T = \rho$ is psd. Thus, Sep $\subseteq$ PPT. On the pset, you will show that if $\rho \in D_{d^2}$ and $\rho^\Gamma \geq 0$, then Tr $[\rho^\Gamma \Phi_d] \leq 1/d$. Let us now consider the composition of the partial transpose and LOCC operations:

**Claim** If $\rho \in$ PPT, and $\mathcal{E}$ is a LOCC operation, then $\mathcal{E}(\rho) \in$ PPT. This result extends to SLOCC (stochastic LOCC).

To quickly recap some definitions, local operations and classical communication channels are described by

$$\rho \to (U \otimes V)\rho(U \otimes V)^\dagger \tag{14.4}$$

Measurement channels,

$$\sum_k (E_k \otimes I)\rho(E_k \otimes I)^\dagger \text{ with } \sum E_k^\dagger E_k \leq \mathbb{I} \tag{14.5}$$

and stochastic LOCC

$$\rho \to (E_k \otimes I)\rho(E_k \otimes I)^\dagger \text{ or } (I \otimes E_k)\rho(I \otimes E_k)^\dagger \tag{14.6}$$

In general,

$$\rho \to (A \otimes B)\rho(A \otimes B)^\dagger \tag{14.7}$$

Where $A, B$ are arbitrary or perhaps invertable.

$$((A \otimes B)\rho(A \otimes B)^\dagger)^\Gamma = (A \otimes \bar{B})\rho^\Gamma(A \otimes \bar{B})^\dagger \geq 0 \qquad \text{if } \rho^\Gamma \geq 0 \tag{14.8}$$

Where is this useful? $D_p$ is never antidegradable for $p < 1$, but is PPT for $p$ large enough.

$\rho \in$PPT $\implies E_{D,2}(\rho) = 0$. PPT = Sep only if $d_A = 2, d_B = 3$.

### 14.3.1 States

As previously argued, there is a straightforward inclusion statement between the set of separable states and PPT

$$\text{Sep} \subset \text{PPT} \subset \text{All} \tag{14.9}$$

Doherty, Parrilo and Spedalieri defined the DPS hierarchy (quant-ph/0308032), based on iteratively extending approximations to the set of entangled states.

$$\text{PPT} = \text{DPS}_1 \supset \text{DPS}_2 \supset \cdots \supset \text{DPS}_\infty = \text{Sep} \tag{14.10}$$

Each $\text{DPS}_k$ requires time $d^{\mathcal{O}(k)}$ to search, so there is a tractable test for entanglement that increases exponentially with $k$.

### 14.3.2 Operations

$$1 - \text{LOCC} \supset \text{LOCC} \supset \text{Sep} \supset \text{PPT} \supset \text{All} \tag{14.11}$$

PPT operations map PPT states onto PPT states. PPT channels always output PPT states.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 15: October 22, 2020

*Lecturer: Aram Harrow* *Scribe: Michael DeMarco, Leon Ding*

## 15.1 Proofs of the quantum capacity formula

1. Coherent Classical Communication and $C_E$
2. Decoupling and merging

The first proof is a simpler one that Aram came up with, but has fewer generalizable insights for quantum information.

Recall that the formal definition of the quantum capacity is:

$$Q = \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \max \left\{ d : d\text{-dimensional subspace } V \text{ of } A^n \text{ s.t. } \forall \left|\psi\right\rangle \in V,\ \mathcal{D}(N^{\otimes n}(\psi)) \approx_\epsilon \psi \right\} \tag{15.1}$$

Remember that $\approx_\epsilon$ means approximately equal with some error proportional to $\epsilon$, the log of a dimension corresponds to a number of qubits, and $\mathcal{D}$ is a decoding map.

### 15.1.1 Detour: Cobits

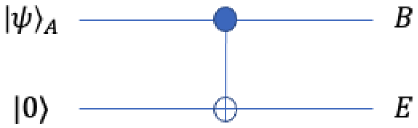
A coherent bit, or cobit (can think of this as intermediate between classical communication and quantum),

$$[q \to q] : a \left|0\right\rangle_A + b \left|1\right\rangle_A \to a \left|0\right\rangle_B + b \left|1\right\rangle_B \tag{15.2}$$

Or more succinctly,

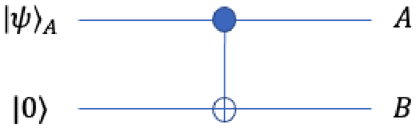
$$\left|x\right\rangle_A \to \left|x\right\rangle_B \text{ for } x \in \{0, 1\}, \text{ isometry} \tag{15.3}$$

Consider classical communication as $[c \to c] : \left|x\right\rangle_A \to \left|x\right\rangle_B \otimes \left|x\right\rangle_E$ using a CNOT gate from $\left|\psi\right\rangle_A \otimes \left|0\right\rangle \to (B, E)$.



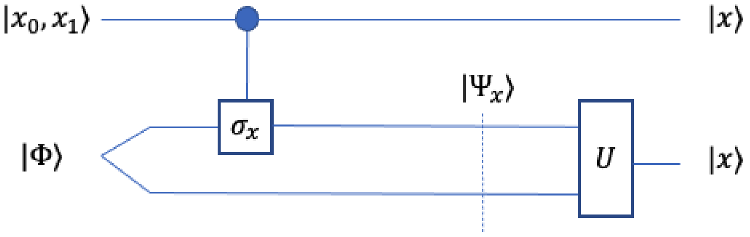
Instead of giving one bit to the environment, what happens if one of the outputs remains on Alice's side?

$$[c \to cc]\,|x\rangle_A \to |x\rangle_A \otimes |x\rangle_B \tag{15.4}$$



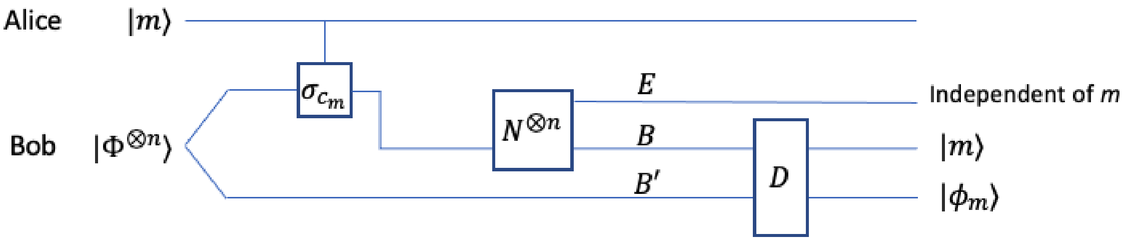
This is the cobit channel. Now, $[q \to q] \geq [c \to cc] \geq [c \to c]$. In fact, we will see that asymptotically $[c \to cc] = \frac{1}{2}([q \to q] + [qq])/2$.

This equality is true because of *decoupling*. Now, cbits (in input or output) do not necessarily leak to E, or at least there is nothing in the environment that leaks to E.



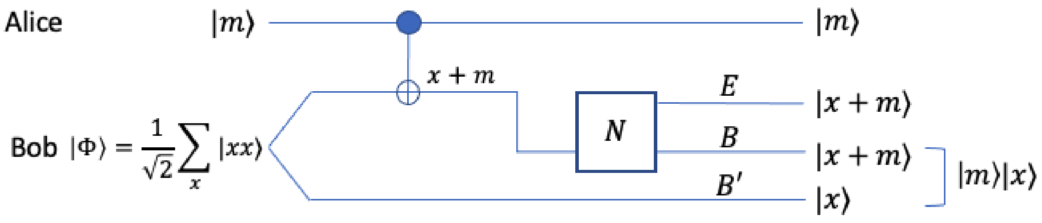
Output Rule (concerning the case with decoupled outputs): super-dense coding $[q \to q] + [qq] \geq 2[c \to c]$ does not leak to environment, hence if we do not throw out bits we get a free upgrade to $[q \to q] + [qq] \geq 2[c \to cc]$. Here, instead of performing a bell-state measurement at the end of the circuit, Bob just applies a unitary $U$ to transform the bell states back to the computational basis.

In general, coherently decoupled $[c \to c]$ (ie where the environment cannot break superpositions of outputs) can turn into $[c \to cc]$.



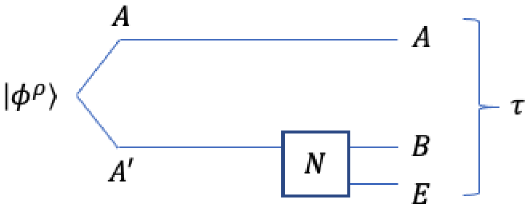
Consider the above example circuit for entanglement assisted communication. Here, Alice and Bob share $n$ copies of the Bell state $|\Phi\rangle$ (of course, it doesn't have to be a Bell state in the general case). Alice performs a controlled Pauli operation depending on some code-word $c_m$, $\mathcal{N}$ is a noisy channel, and $\mathcal{D}$ is Bob's decoder which produces $m$. Bob can erase the content of $|\phi_m\rangle$ using his knowledge of $m$, in which case he is left with a cobit.

Suppose that $N$ is a cbit channel, $m \in \{0, 1\}$. Consider the circuit:



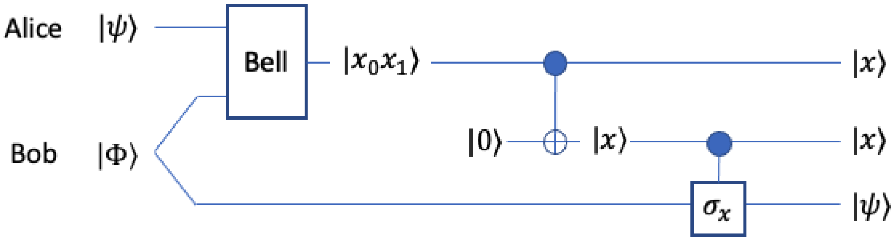
This is a Vernom cipher or one-time pad. Here, Bob may unitarily transform his result into $|m\rangle\,|x\rangle$, without Eve determining the content of $m$. Cobits are, in a sense, the quantum version of the one-time pad.

In fact, the ebit cost is $S(A)$:



and here $\langle N\rangle + S(A)[qq] \geq I(A:B)[c \to cc]$

Input Rule: cobits in decoupled inputs yield ebit teleportation:



First, Alice, instead of doing a Bell measurement, does a unitary transformation from the Bell basis into the standard basis. In normal teleportation, this state would be used to control a controlled-$\sigma_x$ operation, and Bob would be left with the teleported state. Instead of using this classical communication of the bit, we now use a cobit to transmit the information to Bob, with the additional benefit of retaining to two ebits. In resource notation, this process has produced $2[c \to cc] + [qq] \geq 2[qq] + [q \to q]$. However, considering coherent superdense coding $[qq] + [q \to q] \geq 2[c \to cc]$. Hence $2[c \to cc] = [qq] + [q \to qq]$, where the equality holds catalytically. This means that we had one unit of $[qq]$ which was just there as a catalyst.

### 15.1.2 Coherent Classical Communication and $C_E$

Combining this with our earlier observation, we have that $\langle N\rangle + S(A)[qq] \geq \frac{I(A:B)}{2}([q \to q] + [qq])$. This implies:

$$S(A) - \frac{I(A:B)}{2} = \frac{2S(A) - (S(A) + S(B) - S(E))}{2} = \frac{I(A:E)}{2} \tag{15.5}$$

which implies that:

$$\langle N\rangle + \frac{1}{2}I(A:E)[qq] \geq \frac{I(A:B)}{2}[q \to q] \tag{15.6}$$

This is called the 'father' protocol because we can combine it with $[q \to q] \geq [qq]$ (entanglement distribution) to get that:

$$\langle N\rangle \geq \frac{I(A:B) - I(A:E)}{2}[q \to q] = I_C[q \to q] \tag{15.7}$$

where the equality follows from expanding mutual information in terms of entropies. (requires catalytic entanglement use). If we combine the 'father' protocol with superdense coding, we get

$$\langle N\rangle + S(A)[qq] \geq I(A:B)[c \to c] \tag{15.8}$$

Aside: there is also a 'mother' protocol:

$$\langle \rho\rangle + \frac{1}{2}I(A:E)[q \to q] \geq \frac{I(A:B)}{2}[qq] \tag{15.9}$$

which we can combine with teleportation to get:

$$\langle \rho\rangle + I(A:E)[c \to c] \geq I_c(A\rangle B)[qq] \tag{15.10}$$

More details at `quant-ph=0307031` and `quant-ph/03/08/0447`.

### 15.1.3 Decoupling and Merging

Quantum state merging and negative information `quant-ph/0512247` and `quant-ph/0606225`. The 'mother' protocol leads to the mother of all protocols.

The merging task: purify $\rho^{AB}$ to $\psi^{ABR}$. Think of $R$ as a reference system that keeps track of the original message. Goal is for Alice to transmit her half of the state to Bob. Allow free LOCC, ebit cost of merging is $S(A|B)$. If $S(A|B) > 0$, merging is possible by consuming $S(A|B) + \delta$ ebits $\forall \delta > 0$. If $S(A|B) < 0$, then merging is possible while generating $-S(A|B) - \delta$ ebits $\forall \delta > 0$. Either way, we use $I(A : R)$ cbits.

Examples:

1. $\rho = \frac{I}{2}^A \otimes \sigma^B, |\psi\rangle^{ABR} = |\psi\rangle^{AR} \otimes |\phi\rangle^{BR'}$ comm cost is 1.
2. $\rho = \phi^{AB}$, comm cost is $-2$

(Proof to be covered in detail next lecture.)

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 16: October 27, 2020

*Lecturer: Aram Harrow* *Scribe: Joshua Lin, Andrey Boris Khesin*

In this lecture, we will be covering random states and unitaries. One application of random unitaries is that we can use them to destroy information in a certain sense (see the last lecture). A certain theme that we will see repeat is 'concentration of measure', as we take the limit of high dimensions (e.g. if we have large quantum systems), certain natural measures will concentrate along physically interesting subspaces.

### 16.0.4 Scalar Random Variables

**Lemma 8 (Markov's Inequality)**

$$X \geq 0 \implies Pr[X \geq a\mathbb{E}X] \leq \frac{1}{a}$$

To get a tighter bound, use higher moments:

**Lemma 9 (Chebyshev's Inequality)**

$$\mathbb{E}X = \mu,\ \mathbb{E}(X-\mu)^2 = \sigma^2 \implies Pr[(X-\mu)^2 \geq a\sigma] \leq \frac{1}{a^2}$$

This is an example of the "Bernstein trick":

$$f(x) > 0,\ f'(x) \geq 0 \implies \Pr[X \geq a] = \Pr[f(x) \geq f(a)] \leq \frac{E[f(x)]}{f(a)}$$

in other words, to derive Chebyshev's Inequality, we apply Markov's Inequality to a transformed version of the random variable, where we take $f(x)$ to be squared difference from expectation (caveat, $f$ is not monotone here).

Note that Chebyshev's inequality is really weak for things like Gaussian random variables, it tells us that the probably of having a gaussian r.v. $5\sigma$ above average is only bounded by 1/25 (we know that it's very small in reality). Instead, use $f(x) = e^{\lambda x}$ in Bernstein trick, and we get the much better:

$$X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1),\ \Pr[X \geq a] = e^{\lambda^2/2 - \lambda a}$$

This bound is funny because $\lambda$ is a parameter we are free to choose. Clearly, some values of $\lambda$ give better bounds than others, optimal bound at

$$\lambda = a \implies \Pr[X \geq a] \leq e^{-a^2/2}$$

Note that this is a much better bound than just using Chebyshev's Inequality. Actually this bound is pretty much optimal; it turns out it's only off by roughly a constant factor in this particular example - and the only thing we really needed to know is $\mathbb{E}e^{\lambda X}$ called the 'moment generating function', :

$$\mathbb{E}e^{\lambda X} = 1 + \lambda \mathbb{E}X + \frac{\lambda^2}{2}\mathbb{E}X^2 + \dots$$

**Lemma 10 (Chernoff bound)** *$X_1, \dots, X_n$ i.i.d $Pr[X_i = 1] = Pr[X_i = -1] = \frac{1}{2}$, and let $X = \sum_i X_i$.*

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}[e^{\lambda X_1}]^n = (\cosh(\lambda))^n \leq e^{n\lambda^2/2}$$
$$Pr[X \geq \delta n] \leq e^{n\lambda^2/2 - n\delta\lambda} = e^{n\delta^2/2} \qquad if \quad \lambda = \delta$$

We get similar bounds if $|X_i| \leq 1$ and $\mathbb{E}X_i = 0$. Intuitively the same bound still applies, because we will only get less deviation if we allow the random variables to be between $-1$ and 1.

### 16.0.5 Random Vectors

Gaussian random vectors look like:

$$|g\rangle = \begin{pmatrix} g_1 \\ \vdots \\ g_d \end{pmatrix}, \quad g_i \in \mathcal{N}_{\mathbb{C}}\left(0, \frac{1}{d}\right), g_i = x_i + iy_i, \quad x_i, y_i \in \mathcal{N}\left(0, \frac{1}{2d}\right)$$

which gives us $\mathbb{E}\langle g|g\rangle = 1$, and $p(|g\rangle) = \left(\frac{d}{\pi}\right)^d e^{-d\langle g|g\rangle}$. We can use this to generate a random unit vector:

$$|v\rangle = \frac{|g\rangle}{\sqrt{\langle g|g\rangle}}.$$

Gaussian vectors are great because the entries are independent, and the distribution is rotationally symmetric (i.e. it is symmetric under actions of $U(d)$). It turns out that the Gaussian distribution is the only distribution with these properties.

Note that we have $\mathbb{E}[g_i] = 0$ since the distribution of $g_i$ is invariant under phase rotations due to the unitary invariance of the gaussian vector. The only way to get nonzero expectations is to write down 'scalar quantities' under the group invariance:

$$\mathbb{E}[g_i g_j^*] = \frac{\delta_{ij}}{d}, \quad \mathbb{E}[|g\rangle\langle g|] = \sum_{ij} \mathbb{E}[g_i g_j^*]|i\rangle\langle j| = \frac{I}{d}$$

$$\mathbb{E}[g_i g_j g_k^* g_l^*] = \mathbb{E}[g_i g_k^*]\mathbb{E}[g_j g_l^*] + \mathbb{E}[g_i g_l^*]\mathbb{E}[g_j g_k^*] = \frac{\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{d^2}$$

by Isserlis' (/Wick's) theorem - since we know the only way to get nonzero answers out is to pair up the $g$ factors.

$$\mathbb{E}[|g,g\rangle\langle g,g|] = \frac{1}{d^2}\sum_{ij} |i,j\rangle\langle i,j| + |i,j\rangle\langle j,i| = \frac{I + \text{SWAP}}{d^2}$$

where the SWAP operator swaps the two registers. In general, using Wick's theorem:

$$\mathbb{E}[g_{i_1}\dots g_{i_n} g_{j_1}^* \dots g_{j_n}^*] = \frac{1}{d^n}\sum_{\pi\in S_n}\prod_{l=1}^{n} \delta_{i_l, j_{\pi(l)}}$$

$$\mathbb{E}[|g\rangle\langle g|^{\otimes n}] = \frac{1}{d^n}\sum_{\pi\in S_n} P_\pi, \quad P_\pi = \sum_{i_1,\dots,i_n} |i_1,\dots,i_n\rangle\langle i_{\pi(1)},\dots,i_{\pi(n)}|$$

where $S_n$ is the symmetric group, $\pi$ is a permutation, and we are essentially just summing over ways of matching up the $g$ factors. Now, how do we interpret this quantity?

$$\Pi_{\text{sym}} := \frac{1}{n!}\sum_{\pi\in S_n} P_\pi = \text{ projector onto symmetric subspace}$$

$$\text{Sym}^n\mathbb{C}^d = \{|\psi\rangle \in (\mathbb{C}^d)^{\otimes n} : P_\pi|\psi\rangle = |\psi\rangle \quad \forall \pi \in S_n\}$$

Note that we can prove the above fact in much bigger generality, suppose I have an arbitrary finite group $G$ with unitary rep $r : G \to U(V)$, let $V^G$ be the $G$-invariant vectors in $V$, then the claim is that:

$$\Pi = \frac{1}{|G|}\sum_{g\in G} r(g) \quad \text{projects onto} \quad V^G$$

First, note that $\Pi$ itself is invariant under $G$-action:

$$r(h)\Pi = r(h)\frac{1}{|G|}\sum_g r(g) = |G|^{-1}\sum_g r(hg) = |G|^{-1}\sum_g r(g) = \Pi$$

note that group acts freely on itself in the sense that its action is 1-to1. So, that means $r(h)\Pi|\psi\rangle = \Pi|\psi\rangle$, so $\text{Im}(\Pi) \subset V^G$, convsersely $|\psi\rangle \in V^G \implies \Pi|\psi\rangle \in \text{Im}(\Pi)$, finally, to prove that it's a projector:

$$\Pi^\dagger\Pi = \mathbb{E}_h r(h^{-1})\Pi = \Pi$$

so, going back to what we had before,

$$\mathbb{E}[|g\rangle\langle g|^{\otimes n}] = \frac{1}{d^n}\sum_{\pi\in S_n} P_\pi = \frac{n!}{d^n}\Pi_{\text{sym}}$$

Now, what about random unit vectors? Note that $|g\rangle = r|u\rangle$ with $r, |u\rangle$ independent, so that:

$$\mathbb{E}[|g\rangle\langle g|^{\otimes n}] = \mathbb{E}[r^{2n}]\mathbb{E}[|u\rangle\langle u|^{\otimes n}]$$

so, we know in fact that:

$$\mathbb{E}[|u\rangle\langle u|^{\otimes n}] = \frac{\Pi_{\text{sym}}}{\text{tr}\Pi_{\text{sym}}} \implies \text{Sym}^n\mathbb{C}^d = \text{span}\{|\psi\rangle^{\otimes n} : |\psi\rangle \in \mathbb{C}^d\}$$

If we have another basis $p \in P_n$ where $p$ describes a partition, describing a 'type' of unit vector:

$$|p\rangle = \binom{n}{np}^{-1/2} \sum_{x \in T_p^n} |x\rangle$$

$$\mathbb{E}[|u\rangle\langle u|^{\otimes n}] = \frac{\Pi_{\text{sym}}}{\binom{d+n-1}{n}} = \frac{\sum_\pi P_\pi}{d(d+1)\dots(d+n-1)}$$

$$1 \le \mathbb{E}[r^{2n}] = \frac{d(d+1)\dots(d+n-1)}{d^n} \le e^{n^2/2d}$$

So, if $d \gg n$, we have a concentration of measure of the gaussian vectors around the unit vectors.

### 16.0.6 Applications to Entanglement of random states

Suppose that we have a uniformly random $|\psi\rangle \in \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$, then how entangled is $\psi$, in other words, what is $\mathbb{E}S(A)_\psi$?

$$\mathbb{E}S(A)_\psi \ge \mathbb{E}S_2(\psi_A) = -\mathbb{E}\log \text{ tr } \psi_A^2$$

where $S_2(\rho) = -\log \text{tr } \rho^2$ is the Renyi entropy. But, by concavity of log, we have:

$$\mathbb{E}S(A)_\psi \ge -\log \mathbb{E} \text{ tr } \psi_A^2$$

Now, note this cool fact:

$$\text{tr}(X \otimes Y)\text{SWAP} = \text{tr}[XY]$$

which we can draw as a circuit. This gives us:

$$\text{tr}\psi_A^2 = \text{tr}(\psi_A \otimes \psi_A)\text{SWAP} = \text{tr}[(\psi_{A_1B_1} \otimes \psi_{A_2B_2})(\text{SWAP}_{A_1A_2} \otimes I_{B_1B_2})]$$

But now, we have that:

$$\mathbb{E}\text{tr}(\psi_A^2) = \text{tr}\Big[\mathbb{E}(\psi_{A_1B_1} \otimes \psi_{A_2B_2})\,(\text{SWAP}_{A_1A_2} \otimes I_{B_1B_2})\Big]$$

$$= \operatorname{tr}\frac{\text{SWAP}_{A_1A_2} + \text{SWAP}_{B_1B_2}}{d_Ad_B(d_Ad_B+1)} = \frac{d_Ad_B^2 + d_A^2d_B}{d_Ad_B(d_Ad_B+1)} = \frac{d_A+d_B}{d_Ad_B+1}$$

So, putting all this together, we get that

$$\mathbb{E}S(A)_\psi \geq \log\left(\frac{d_Ad_B+1}{d_A+d_B}\right)$$

So, if the dimensions are equal, our bound is roughly $\log(d)-1$, and if we have $d_A \ll d_B$, then our bound looks instead like $\log(d_A) - \frac{d_A}{d_B}$ (which means you have a small correction to that of the maximally entangled state).

The takeaway is that random states are close to maximally entangled, with small corrections due to the finiteness of the dimensions.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 17: October 29, 2020

*Lecturer: Aram Harrow* *Scribe: Annie Wei*

## 17.1 Entanglement of random states

Recall that last time we studied entanglement in random states. We showed that for $|\psi\rangle \in \mathbb{C}^{d_a} \otimes \mathbb{C}^{d_b}$,

$$\begin{aligned} \mathbb{E} S(A)_\psi &\geq \mathbb{E} S_2(\psi_A) \qquad (17.1) \\ &\geq -\log \mathbb{E}\, \mathrm{tr} \psi_A^2 \qquad (17.2) \end{aligned}$$

where

$$\mathbb{E}\, \mathrm{tr} \psi_A^2 = \frac{d_A + d_B}{d_A d_B + 1} \tag{17.3}$$

For $d = d_A = d_B$, we get

$$\mathbb{E} S(A)_\psi \geq \log \frac{d^2+1}{2d} \geq \log(d) - 1 \tag{17.4}$$

For $d_A << d_B$, we get

$$\mathbb{E} S(A)_\psi \geq \log(d_A) - \log(1 + \frac{d_A}{d_B}) \approx \log d_A \tag{17.5}$$

That is, a random $n$-qubit state has $k$-qubit marginals that look like $I/2^k$ if $k < n/2$.

How accurate is this bound? Suppose that

$$|\psi\rangle = \sum_{ij} G_{ij} |i\rangle \otimes |j\rangle \tag{17.6}$$

with

$$\mathbb{E} |G_{ij}|^2 = \frac{1}{d_A d_B} \tag{17.7}$$

Then this has marginals $\psi_A = GG^\dagger$, corresponding to the complex Wishart distribution.

The histogram of eigenvalues $\lambda$ of $\psi_A$ follows the Marchenko-Pastor laws and satisfies

$$\lambda_{min} \approx \frac{1}{d_A}\left(1 - \sqrt{\frac{d_A}{d_B}}\right)^2 \tag{17.8}$$

$$\lambda_{max} \approx \frac{1}{d_A}\left(1 + \sqrt{\frac{d_A}{d_B}}\right)^2 \tag{17.9}$$

$$\mu(\lambda) = \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{2\pi(d_A/d_B)\lambda} \tag{17.10}$$

where $\mu(\lambda)$ is the density. A sketch of the proof goes like the following:

$$\mathrm{tr}\psi_A^k = \binom{d_A + k - 1}{k}^{-1} \mathrm{tr}\Pi_{sym}(C_{A_1,\dots,A_k} \otimes I_{B_1,\dots,B_k}) \tag{17.11}$$

$$\approx d_A \int \mu(\lambda)\lambda^k d\lambda \tag{17.12}$$

A special case of this is when $d = d_A = d_B$. Then $\lambda_{min} \sim 1/d^2$, $\lambda_{max} \approx 4/d$, and

$$\mu(\lambda) = \frac{\sqrt{4/d - \lambda}}{2\pi\sqrt{\lambda}} \tag{17.13}$$

$$\mu(\sqrt{\lambda}) = \frac{\sqrt{4/d - \lambda}}{\pi} \tag{17.14}$$

This is known as the quarter-circle law (note that there's a Wigner semicircle law for eigenvalues of $G + G^\dagger$, and a circle law for eigenvalues of $G$).

## 17.2 Note on Renyi entropies

Suppose we have a state

$$\rho = \frac{1}{2}\left|0\right\rangle\left\langle 0\right| \otimes (I/2)^{\otimes a} \otimes \left|0\right\rangle\left\langle 0\right|^{\otimes(b-a)} + \frac{1}{2}\left|1\right\rangle\left\langle 1\right| \otimes (I/2)^{\otimes b} \tag{17.1}$$

for $a < b$. Then

$$S_0(\rho) = \log(2^a + 2^b) \approx b + 2^{a-b} \tag{17.2}$$

$$S_\infty(\rho) = a + 1 \tag{17.3}$$

$$S(\rho) = 1 + \frac{a+b}{2} \tag{17.4}$$

$$S_\alpha(\rho) = \frac{1}{1-\alpha}\log(2^a(2^{a+1})^{-\alpha} + 2^b(2^{b+1})^{-\alpha}) \tag{17.5}$$

$$= \frac{1}{1-\alpha}[\log((2^{1-\alpha})^a + (2^{1-\alpha})^b) - \alpha] \tag{17.6}$$

For $\alpha > 1$ the first term dominates, while for $\alpha < 1$ the second term dominates. This is why we like taking $\alpha = 1$, where the contributions are the same

## 17.3 k-designs

Say that $\mu$ is a distribution on $S_d$, the states in $\mathbb{C}^d$. Then $\mu$ is a k-design if

$$\mathbb{E}_{|\psi\rangle\sim\mu}\psi^{\otimes k} = \mathbb{E}_{\psi\sim\text{Uniform}}\psi^{\otimes k} = \Pi_{sym}\binom{d+k-1}{k}^{-1} \tag{17.1}$$

Note that we can also define approximate k-designs. 1-designs are pretty easy to come up with, i.e. $\{|000\rangle, ..., |111\rangle\}$ is a 1-design. Stabilizer states are 2-designs (and also 3-designs). (Recall that stabilizer states are those that can be written as $C|0^n\rangle$ for $C$ a Clifford state. Alternatively, we can define them as the simultaneous $+1$ eigenstate of $n$ commuting operators of the form $\sigma_{i_1}\otimes\sigma_{i_2}\otimes....\otimes\sigma i_n$.)

### 17.3.1 Application: $\epsilon$-randomizing maps

We say that $N : D_d \to D_d$ is $\epsilon$-randomizing if $\forall\ \rho$,

$$||N(\rho) - I/d||_\infty \leq \epsilon/d \tag{17.2}$$

We will consider maps of the form

$$N(\rho) = \frac{1}{n}\sum_{i=1}^{n} U_i\rho U_i^\dagger \tag{17.3}$$

How large does $n$ need to be? (Recall that we can do remote state preparation with cost $\log n$.) Note that

$$\operatorname{rank} N(|1\rangle\langle 1|) \leq n \tag{17.4}$$

$$\tag{17.5}$$

For a choice of $\epsilon < 1$,

$$||N(|1\rangle\langle 1|) - I/d||_\infty \leq 1/d \tag{17.6}$$

$$\Rightarrow \operatorname{rank} N(|1\rangle\langle 1|) = d \tag{17.7}$$

Thus $n \geq d$.

Note that the generalized Paulis work with $n = d^2$, $\epsilon = 0$. In fact, $\epsilon = 0$ allows for teleportation. To see this, note that $\epsilon = 0 \Rightarrow N(\rho) = I/d \Rightarrow N(X) = \operatorname{tr}(X)I/d$ by linearity. Then

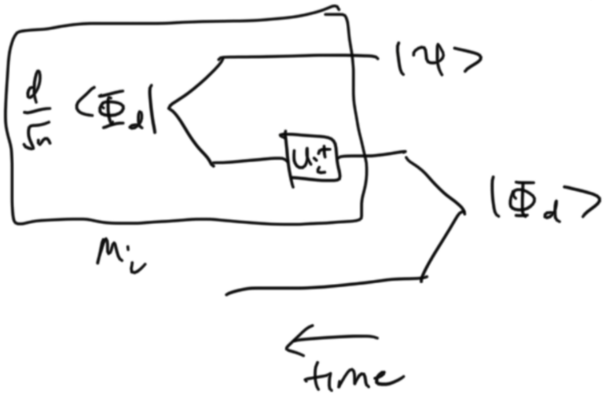
$$(I\otimes N)(\Phi_d) = \frac{1}{d}\sum_{ij}|i\rangle\langle j|\otimes N(|i\rangle\langle j|) \tag{17.8}$$

$$= (I/d)\otimes(I/d) \tag{17.9}$$

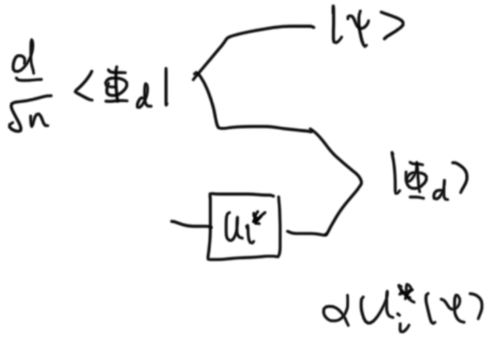
Thus the set of operators

$$M_i = \frac{d^2}{n}(I \otimes U_i)\Phi_d(I \otimes U_i^\dagger) \tag{17.10}$$

is PSD and satisfies $\sum_i M_i = I$, so it forms a POVM. We can draw the following diagram for a teleportation protocol:



Moving the unitary and transposing, this becomes


$\frac{d}{\sqrt{n}}\langle\Phi_d|$
$|\psi\rangle$
$|\Phi_d\rangle$
$U_i^*$
$\propto U_i^*|\psi\rangle$

Note that this also gives us a lower bound $n \geq d^2$ (otherwise we could do teleportation with less than $n < d^2$).

If we let $\epsilon > 0$, it's possible to have $n = O(d/\epsilon^2)$. Let

$$\alpha = \max_{\rho} ||N(\rho) - I/d||_\infty = \epsilon/d \tag{17.11}$$

$$= \max_{\rho,\sigma} |\mathrm{tr}(N(\rho) - I/d)\sigma| \tag{17.12}$$

$$= \max_{|\rho\rangle,|\varphi\rangle} |\mathrm{tr}N(\psi)\varphi - 1/d| \tag{17.13}$$

$$= \max_{|\rho\rangle,|\varphi\rangle} \left| \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}[U_i\psi U_i^\dagger\varphi] - 1/d \right| \tag{17.14}$$

We will later use the fact that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}[U_iAU_i^\dagger B]-1/d\right|\leq ||A||_1||B||_1\left(\frac{1}{d}+\alpha\right) \tag{17.15}$$

for $A$, $B$ Hermitian. Now fix $\psi,\varphi,i$, and let

$$\mathrm{tr}U_i\psi U_i^\dagger\varphi=|\gamma_1|^2 \tag{17.16}$$

where $U_i\left|\psi\right\rangle=\left|\gamma\right\rangle$, $\left|\varphi\right\rangle=\left|1\right\rangle$. Also let $\left|g\right\rangle=r\left|\gamma\right\rangle$ so that

$$\begin{aligned}\mathbb{E}\exp(\lambda|\gamma_1|^2)&\leq\mathbb{E}e^{\lambda r^2}\mathbb{E}\exp(\lambda|\gamma_1|^2) && (17.17)\\ &=\mathbb{E}e^{-\lambda|g_1|^2} && (17.18)\\ &=\frac{1}{1-\lambda/d} && (17.19)\\ \mathbb{E}\exp(\lambda\frac{1}{n}\sum_i\mathrm{tr}[U_i\psi U_i^\dagger\varphi])&\leq\left(1-\frac{\lambda}{nd}\right)^{-n} && (17.20)\end{aligned}$$

after some algebra (see quant-ph/0307100 for more details).

Now, for fixed $\psi,\varphi$, we have that

$$\Pr\left[\left|\frac{1}{n}\sum_i\mathrm{tr}[U_i\psi U_i^\dagger\varphi]-\frac{1}{d}\right|\geq\epsilon/d\right]\leq\exp(-cn\epsilon^2) \tag{17.21}$$

We want to be able to make a statement about

$$\Pr\left[\exists\psi,\varphi\left|\frac{1}{n}\sum_i\mathrm{tr}[U_i\psi U_i^\dagger\varphi]-\frac{1}{d}\right|\geq\epsilon/d\right] \tag{17.22}$$

Normally we would use a union bound, but in this case we need to use a $\delta$-net. Specifically, we say that $M$ is a $\delta$-net if $\forall\left|x\right\rangle\in S_d$, $\exists\left|\beta\right\rangle\in M$ such that $||\left|\alpha\right\rangle-\left|\beta\right\rangle||_2\leq\delta$.

We claim that there exists a $M$ of size $|M|\leq(1+(2/\delta))^{2d}$. To prove this, we add $\left|\beta_1\right\rangle,\left|\beta_2\right\rangle,....$ to $M$ until $||\left|\beta_i\right\rangle-\left|\beta_j\right\rangle||_2>\delta$ no longer holds. Note that the $B(\left|\beta_i\right\rangle,\delta/2)$ are all disjoint and are contained in $B(0,1+\delta/2)$. Letting $\mathrm{Vol}(B(0,r))=C_dr^{2d}$, $|M|C_d(\delta/2)^{2d}\leq C_d(1+\delta/2)^{2d}\Rightarrow|M|\leq(1+2/\delta)^{2d}$.

Now converting this to the trace norm,

$$||\left|\psi\right\rangle-\left|\varphi\right\rangle||_{\ell_2}\geq\frac{1}{2}||\psi-\varphi||_{S_1} \tag{17.23}$$

Thus $M$ is a $\delta$-net with $|M|\leq(3/\delta)^{2d}$.

Now let

$$\beta = \max_{|\psi_0\rangle, |\varphi_0\rangle \in M} \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr} U_i \psi_0 U_i^\dagger \varphi_0 - \frac{1}{d} \right| \tag{17.24}$$

Note that

$$\Pr[\beta \geq \epsilon/d] \leq (3/\delta)^{4d} e^{-cn\epsilon^2} < 1 \tag{17.25}$$

if we choose $\delta = O(1)$, $n = O(d/\epsilon^2)$. Now we just need to extend to points not in the net. Letting

$$||\psi - \psi_0||_1 \leq 2\delta \tag{17.26}$$

$$||\varphi - \varphi_0||_1 \leq 2\delta \tag{17.27}$$

for some $\psi, \varphi$,

$$\alpha = \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr} U_i \psi U_i^\dagger \varphi - \frac{1}{d} \right| \tag{17.28}$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr} U_i \psi_0 U_i^\dagger \varphi_0 - \frac{1}{d} \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr} U_i (\psi - \psi_0) U_i^\dagger \varphi_0 \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{tr} U_i \psi_0 U_i^\dagger (\varphi - \varphi_0) \right| \tag{17.29}$$

$$\leq \beta + 2 \cdot 2\delta \left( \frac{1}{d} + \alpha \right) \tag{17.30}$$

$$\Rightarrow \alpha \leq \frac{1}{1 - 4\delta} (\beta + \frac{4\delta}{d}) = O(\epsilon/\delta) \tag{17.31}$$

Note that $(I \otimes N)\Phi_d$ has rank $d/\epsilon^2$ but is LOCC-indistinguishable from $I/d \otimes I/d$ with rank $d^2$. Thus it accomplishes data hiding.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 18: November 3, 2020

*Lecturer: Aram Harrow* *Scribe: Yogeshwar Velingker, Joshua Lin*

Recall from last time that

$$\mathop{\mathbb{E}}_{|v\rangle \in \mathbb{C}^d} |v\rangle\langle v|^{\otimes n} = \frac{\Pi_{sym}^{d,n}}{\mathrm{tr}\Pi_{sym}^{d,n}} = \frac{\sum_{\pi \in S_n} P_\pi}{d(d+1)\cdots(d+n-1)}.$$

Previously we proved this using direct calculation, but it can be shown in a more illuminating way using representation theory.

## 18.1 Representation theory

Let $G$ be a group, and $V$ a vector space. Then a map $r : G \to L(V)$ is a representation if

$$r(gh) = r(g)r(h)$$

for all $g, h \in G$.

Examples: $g \in U_d \to g^{\otimes n} \in U(\mathbb{C}^{d^n})$

$\pi \in S_n \to P_\pi \in U(\mathbb{C}^{d^n})$ (acts by permuting the positions).

Fix $\omega \in \mathbb{C}$ where $|\omega| = 1$. Then $z \in \mathbb{Z} \to \omega^z \in U(1)$ is a representation. This also works for any $\omega \in \mathbb{C}$, but we may get a non-unitary representation. Similarly, if $\omega^p = 1$, then $z \in \mathbb{Z}_p \to \omega^z \in U(1)$ is a representation.

Two representations $(r_1, V_1)$ and $(r_2, V_2)$ are considered *equivalent* if there exists $T \in L(V_1, V_2)$ such that

$$Tr_1(g) = r_2(g)T$$

for all $g \in G$. Written as $(r_1, V_1) \cong (r_2, V_2)$.

Fact: If $G$ is finite or compact then any representation is equivalent to a unitary representation.

A representation $(r, V)$ is *reducible* if $(r, V) \cong (r_1 \oplus r_2, V_1 \oplus V_2)$. There is a basis in which for all $g$, $r(g)$ can be written in block diagonal form $\left(\begin{array}{c|c} r_1(g) & 0 \\ \hline 0 & r_2(g) \end{array}\right)$. If there is no such decomposition, $(r, V)$ is an *irreducible* representation, or irrep.

Examples: Trivial representation $r(g) = 1 \in U(1)$ for all $g$.

For a finite group $G$, let $\mathbb{C}[G] = \text{span}\{|g\rangle : g \in \mathbb{C}\} \cong \{f : G \to \mathbb{C}\}$. Then we obtain the left/right regular representations: $L(x)|g\rangle = |xg\rangle$ and $R(x)|g\rangle = |gx^{-1}\rangle$. These representations are reducible, since $\sum_{g\in G}|g\rangle$ is acted on trivially.

In fact, we can decompose $\mathbb{C}[G]$ into irreps. Let $\hat{G}$ be the set of inequivalent irreps $(r_\lambda, V_\lambda)$. Then we can write

$$\mathbb{C}[G] \underset{L(g_1)R(g_2)}{\cong} \bigoplus_{\lambda\in\hat{G}} V_\lambda \otimes V_\lambda^* \underset{L}{\cong} \bigoplus_{\lambda\in\hat{G}} V_\lambda \otimes \mathbb{C}^{\dim V_\lambda}$$

where the dual representation $(r^*, V^*)$ is defined as $r^*(g) = r(g^{-1})^T$ and the left representation and right representation act on different spaces. Note that the dimensions of both sides are equal: $|G| = \sum_\lambda d_\lambda^2$ where $d_\lambda$ is the dimension of $V_\lambda$.

We can write $L(V, W) \cong V^* \otimes W$ since linear maps look like $\sum_{v,w} c_{v,w} |w\rangle \langle v|$. If we have two representations $(r, V)$ and $(s, W)$ we obtain a representation $r(g^{-1})^T \otimes s(g) = r^*(g) \otimes s(g)$ acting on matrices as follows:

$$M \in L(V, W) \to s(g)Mr(g)^{-1}$$

since $vec(AMB) = (A \otimes B^T)vec(M)$.

Examples: For unitaries $U$, $r(U) = U \otimes U^*$ corresponds to $\rho \to U\rho U^\dagger$. A 1D invariant subspace is spanned by the maximally entangled state $|\Phi\rangle = \sum_i |i\rangle \otimes |i\rangle = vec(I)$. The remaining $(d^2 - 1)$-dimensional subspace also turns out to be irreducible.

$r(U) = U \otimes U$. This commutes with $F = \text{SWAP}$, so it preserves the $V_{sym}$ and $V_{anti}$ subspaces, which have dimensions $d(d\pm1)/2$. These subrepresentations are irreducible. For $d = 2$, these are known as the triplet and singlet states. In general $(r(U) = U^{\otimes n}, \text{Sym}^n\mathbb{C}^d)$ is an irrep of $U_d$.

Proof sketch: Suppose $|\psi_1\rangle, |\psi_2\rangle \in \text{Sym}^n\mathbb{C}^d$. There exist $|\phi_1\rangle, |\phi_2\rangle$ such that $\langle\psi_1|^{\otimes n}|\psi_i\rangle \neq 0$. This can be used to show the existence of $U$ such that $\langle\psi_1| r(U) |\psi_2\rangle \neq 0$.

Schur's Lemma: If $V_\mu, V_\nu$ are irreps of a group $G$ over $\mathbb{C}$, and then the set of $G$-invariant maps from $V_\mu \to V_\nu$ is

$$L(V_\mu, V_\nu)^G = \begin{cases} 0, & \mu \neq \nu \\ \mathbb{C}I, & \mu = \nu \end{cases}$$

This is the set of maps that preserve the group action, i.e. $r_\nu(g)T = Tr_\mu(g)$ for all $g$.

Proof: Suppose $T \in L(V_\mu, V_\nu)^G$. Then the subspaces $\ker T$ and $\operatorname{Im} T$ are $G$-invariant. Since $V_\mu$ and $V_\nu$ are irreps, either $\ker T = 0$ or $\ker T = V_\mu$, and either $\operatorname{Im} T = 0$ or $\operatorname{Im} T = V\nu$. Therefore, if $\mu \neq \nu$ we must have $T = 0$. Otherwise, if $\mu = \nu$, choose eigenvalue $\lambda$ of $T$ (which exists since it is over $\mathbb{C}$). Then ker $(T - \lambda I) \neq 0$, so ker $(T - \lambda I) = V\mu$ and $T = \lambda I$.

This does not work for irreps over $\mathbb{R}$. Consider the $SO(2)$ action on $\mathbb{R}^2$, and let $T = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. This commutes with the group action, but is not a multiple of $I$.

Before we move on, we will introduce the Haar measure. This is the uniform measure on compact groups, and is the unique measure satisfying

$$\mu_{\text{Haar}}(S) = \mu_{\text{Haar}}(gS) = \mu_{\text{Haar}}(Sg).$$

If $U \sim$ Haar, then for arbitrary $|v\rangle$, $U|v\rangle$ is uniformly random.

$$\mathop{\mathbb{E}}_{U \sim \text{Haar}} r(U)$$

is the projector onto $U_d$-invariant vectors.

Calculation:

Let

$$M = \mathop{\mathbb{E}}_{|\psi\rangle \in \mathbb{C}^d, \langle\psi|\psi=1\rangle} |\psi\rangle\langle\psi|^{\otimes n} = \mathop{\mathbb{E}}_{U\sim\text{Haar}} (U|0\rangle\langle 0| U^\dagger)^{\otimes n}) = \mathbb{E}\, r(U)\, |0\rangle\langle 0|^{\otimes n}\, r(U)^\dagger.$$

Then

$$\begin{aligned} r(V)M &= \mathop{\mathbb{E}}_{U} r(VU)\,|0\rangle\langle 0|^{\otimes n}\, r(U^\dagger) && (18.1)\\ &= \mathop{\mathbb{E}}_{W=VU} r(w)\,|0\rangle\langle 0|^{\otimes n}\, r(W \dagger V) && (18.2)\\ &= Mr(V), && (18.3) \end{aligned}$$

so by Schur's Lemma

$$M = \lambda I_{\text{Sym}^n \mathbb{C}^d} = \frac{\Pi^{d,n}_{\text{Sym}}}{\text{tr}\Pi^{d,n}_{\text{Sym}}}.$$

What about $\mathbb{E}\, U^{\otimes n} M (U^\dagger)^{\otimes n}$ if $M \neq |\psi\rangle\langle\psi|^{\otimes n}$?

Isotypic decomposition: Compact/finite groups satisfy complete reducibility, which means that every representation can be decomposed into a direct sum of irreps.

$$\begin{aligned} V &\cong \bigoplus_{\lambda \in \hat{G}} V_\lambda \otimes \mathbb{C}^{M_\lambda} && (18.4)\\ &\cong \bigoplus_{\lambda} V_\lambda \otimes L(V_\lambda, V)^G && (18.5) \end{aligned}$$

where $M_\lambda \geq 0$ is the multiplicity of $\lambda$.

## 18.2 Schur-Weyl duality

Duality between action of the unitary group and the symmetric group.

Let $q_n(U) = U^{\otimes n}$ and $p_d(\pi) = \sum_{i_1 \dots i_n \in [d]} |i_1 \dots i_n\rangle \langle i_{\pi(1)} \dots i_{\pi(n)}|$, where both operators act on $(\mathbb{C}^d)^{\otimes n}$. We have $[q_n(U), p_d(\pi)] = 0$.

We can write

$$(\mathbb{C}^d)^{\otimes n} \cong^{U_d \times S_n}_{p_d q_n} \bigoplus_{\lambda \in \mathrm{Par}(n,d)} Q_\lambda \otimes P_\lambda$$

where $Q_\lambda$ and $P_\lambda$ are $U_d$ and $S_n$ irreps respectively, labeled by partitions from the set

$$\mathrm{Par}(n,d) = \{\lambda \in \mathbb{Z}^d : \lambda_1 \geq \dots \geq \lambda_d \geq 0, \sum_i \lambda_i = n\}.$$

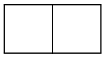
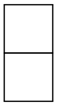
Schur-Weyl duality says that the each $\lambda$ has multiplicity 1.

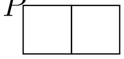
Follows from the following. Let

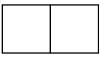
$$A = \mathrm{span}\{q_n(U) : U \in U_d\}, B = \mathrm{span}\{p_d(\pi) : \pi \in S_n\}$$

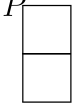
Then $\mathrm{Comm}(A) = \{X : [X, a] = 0 \ \forall a \in A\} = B$ and $\mathrm{Comm}(B) = A = \mathrm{span}\{X^{\otimes n} : X \in M_d\}$ by the Double Commutant theorem.

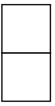
Examples:

We can denote a partition by a *Young diagram*, where we arrange $n$ boxes in rows corresponding to each part. If $n = 2$, there are two partitions: $\lambda = (2, 0)$  and $\lambda = (1, 1)$  and we have

$P$  = trivial

$Q$  = symmetric

$P$  = sign

$Q$  = antisymmetric

If $d = 2$, then for a partition $\lambda = (\lambda_1, \lambda_2)$ we obtain a spin-$J$ representation, where $J = (\lambda_1 - \lambda_2)/2$.

When $n = 3, d = 2$ there are two partitions: $P_{\ydiagram{3}}$ = trivial, $Q_{\ydiagram{3}} =$ spin $3/2, \dim 4$ and $\dim P_{\ydiagram{2,1}} = 2, Q_{\ydiagram{2,1}} =$ spin $1/2, \dim 2$. The dimensions match up since $2^3 = 1 \cdot 4 + 2 \cdot 2$.

#Par $\sim n^d$, $\dim Q_\lambda \leq n^{d^2}$, $\dim P_\lambda \approx \exp(nH(\frac{\lambda}{n}))$.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 19: November 5, 2020

*Lecturer: Aram Harrow* *Scribe: Shreya Vardhan, Yogeshwar Velingker*

## 19.1 Schur-Weyl Duality

Recall from the previous lecture that $(\mathbb{C}^d)^{\otimes n}$ can be decomposed into irreps of $U^d \times S_n$ as:

$$(\mathbb{C}^d)^{\otimes n} \simeq \oplus_{\lambda \in \mathrm{Par}(n,d)}\, Q_\lambda \otimes P_\lambda \tag{19.1}$$

where $\mathrm{Par}(n,d)$ is the set of partitions of $n$ into $d$ elements, each $Q_\lambda$ is an irrep of $V_d$, and each $P_\lambda$ is an irrep of $S_n$.

In particular, for the $n = 2$ case, we have two possible partitions $\lambda = (2,0)$, represented by $\square\square$, and $\lambda = (1,1)$, represented by $\begin{smallmatrix}\square\\\square\end{smallmatrix}$. Corresponding to these, we get two terms in (19.1):

$$(\mathbb{C}^d)^{\otimes 2} \simeq Q_{\square\square} \otimes P_{\square\square} \;\oplus\; Q_{\begin{smallmatrix}\square\\\square\end{smallmatrix}} \otimes P_{\begin{smallmatrix}\square\\\square\end{smallmatrix}} \tag{19.2}$$

where $Q_{\square\square}$ is the $d(d+1)/2$-dimensional symmetric representation of $V_d$, $P_{\square\square}$ is the 1-dimensional trivial representation of $S_2$, $Q_{\begin{smallmatrix}\square\\\square\end{smallmatrix}}$ is the $d(d-1)/2$-dimensional antisymmetric representation of $V_d$, $P_{\square\square}$ is the 1-dimensional sign representation of $S_2$.

## 19.2 Application to merging

Recall from the last lecture that as a consequence of (19.2),

$$\begin{aligned}\mathbb{E}_U\ (U \otimes U)\ X\ (U \otimes U)^\dagger &= \text{projection of X onto } (\mathrm{span}\{\Pi_{\mathrm{sym}}, \Pi_{\mathrm{anti}}\} = \mathrm{span}\{I, F\}) \\ &= \frac{\mathrm{Tr}[X\Pi_{\mathrm{sym}}]}{\mathrm{Tr}\Pi_{\mathrm{sym}}}\Pi_{\mathrm{sym}} + \frac{\mathrm{Tr}\,[X\Pi_{\mathrm{anti}}]}{\mathrm{Tr}\,\Pi_{\mathrm{anti}}}\Pi_{\mathrm{anti}}\end{aligned} \tag{19.1}$$

This tells us that $\mathbb{E}\ (U \otimes U)\ X\ (U \otimes U)^\dagger$ has a simple block-diagonal structure when written in terms of the symmetric and antisymmetric subspaces, and is proportional to the identity within each block.

We consider the particular case of $X = \psi \otimes \psi$, and consider the setup for decoupling, shown in figure 19.1.

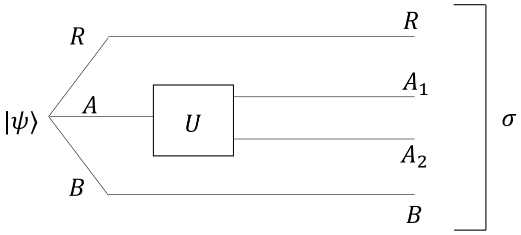


Figure 19.1: Setup for decoupling

Decoupling implies that $\sigma^{A_2R} \approx \sigma^{A_2} \otimes \sigma^R$. Let us see this by first looking at the distance in the Schatten 2-norm, which is mathematically easier to work with despite being less useful operationally:

$$\begin{aligned}\mathbb{E}_U||\sigma_{A_2R} - \sigma_{A_2} \otimes \sigma_R||_2^2 \\ = \mathbb{E}_U[\mathrm{tr}\sigma_{A_2R}^2 - 2\mathrm{tr}(\sigma_{A_2R}\sigma_{A_2} \otimes \sigma_R) + \mathrm{tr}\sigma_{A_2}^2\mathrm{tr}\sigma_R^2]\end{aligned} \tag{19.2}$$

Let us now compute each of the terms in (19.2). The first term can be evaluated as follows:

$$\begin{aligned}\mathbb{E}\mathrm{tr}\sigma_{A_2R}^2 = \mathbb{E}\mathrm{tr}(\sigma_{A_2R} \otimes \sigma_{A_2'R'})F^{A_2R} \\ = \mathrm{tr}[\psi_{AR} \otimes \psi_{A'R'}\mathbb{E}(U^\dagger \otimes U^\dagger)F^{A_2R}(U \otimes U)]\end{aligned} \tag{19.3}$$

where for instance $A'$ refers to a copy of the system $A$, and $F_B$ for any subsystem $B$ refers to the swap operator on two copies of $B$. $F$ has a non-trivial evolution only in $A_2$, so we find

$$\mathbb{E}(U^\dagger \otimes U^\dagger)F^{A_2}(U \otimes U) = \alpha_+\frac{\Pi_+}{\mathrm{Tr}[\Pi_+]} + \alpha_-\frac{\Pi_-}{\mathrm{Tr}[\Pi_-]}, \quad \Pi_\pm = \frac{I \pm F}{2} \tag{19.4}$$

where

$$\alpha_\pm = \mathrm{tr}[\Pi_\pm F^{A_2}] = \mathrm{tr}[\frac{I \pm F^{A_1}F^{A_2}}{2}\ F^{A_2}] = \mathrm{tr}\frac{[F^{A_2} \pm F^{A_1}]}{2} \tag{19.5}$$

so that

$$\begin{aligned}\mathbb{E}_{U_{A_1A_2}}(U^\dagger \otimes U^\dagger)F^{A_2}(U \otimes U) = \frac{d_{A_1} + d_{A_2}}{d_A + 1}\Pi_+ + \frac{d_{A_1} - d_{A_2}}{d_A - 1}\Pi_- \\ \equiv p\ \Pi_+ + q\ \Pi_- \\ = \frac{p+q}{2}I + \frac{p-q}{2}F^A\end{aligned} \tag{19.6}$$

So overall,

$$\begin{aligned}\mathbb{E}\ \mathrm{tr}[\sigma_{A_2R}^2] = \mathrm{tr}[(\psi_{AR} \otimes \psi_{A'R'})(\frac{p+q}{2}I + \frac{p-q}{2}F^A)F^R \\ = \frac{p+q}{2}\mathrm{tr}\psi_R^2 + \frac{p-q}{2}\mathrm{tr}\psi_{AR}^2\end{aligned} \tag{19.7}$$

Note that

$$\frac{p+q}{2} \approx d_{A_2}^{-1}, \quad \frac{p-q}{2} \approx d_{A_1}^{-1} \tag{19.8}$$

The last term in (19.2) is

$$\mathbb{E}[\mathrm{tr}\sigma_{A_2}^2 \mathrm{tr}\sigma_R^2] = \mathbb{E}[\mathrm{tr}\,\sigma_{A_2}^2]\ \mathrm{tr}\,\psi_R^2, \tag{19.9}$$

where

$$\begin{aligned}\mathbb{E}\mathrm{tr}\sigma_{A_2}^2 &= \mathbb{E}\,\mathrm{tr}(U_A \otimes U_{A'})(\psi_A \otimes \psi_{A'})(U_A \otimes U_{A'})^\dagger F^{A_2} \\ &= \mathrm{tr}(\psi_A \otimes \psi_{A'})(\frac{p+q}{2} I + \frac{p-q}{2} F^A) \\ &= \frac{p+q}{2} + \frac{p-q}{2}\mathrm{tr}(\psi_A^2) \leq p = \frac{d_A + d_{A_2}}{d_{A_1} d_{A_2} + 1} \approx \frac{1}{d_{A_2}}\end{aligned} \tag{19.10}$$

if $d_{A_1} > d_{A_2}$. This suggests $\sigma_{A_2}$ is close to the maximally mixed state if $d_{A_1} > d_{A_2}$.

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\sigma_{A_2R}(\sigma_{A_2} \otimes \sigma_R) &= \mathbb{E}\,\mathrm{tr}(U_A \otimes U_{A'})(\psi_{AR})(U_A \otimes U_{A'})^\dagger F^{AR} \\ &= \mathrm{tr}(\psi_{AR} \otimes \psi_{A'} \otimes \psi_{R'})(\frac{p+q}{2} I + \frac{p-q}{2} F^A) F^R \\ &= \frac{p+q}{2}\mathrm{tr}\psi_R^2 + \frac{p-q}{2}\mathrm{tr}\psi_{AR}(\psi_A \otimes \psi_R)\end{aligned} \tag{19.11}$$

Putting all the terms together,

$$\mathbb{E}\,||\sigma_{A_2R_2} - \sigma_{A_2} \otimes \sigma_R||_2^2 = \frac{d_{A_1}(d_{A_2}^2 - 1)}{d_A^2 - 1}(\mathrm{tr}\psi_{AR}^2 - 2\mathrm{tr}\psi_{AR}(\psi_A \otimes \psi_R) + \mathrm{tr}\psi_A^2 \mathrm{tr}\psi_R^2) \tag{19.12}$$

Note that we have made no approximations so far, and this expression is exact. Moreover, we only used the fact that $U$ is a “2-design.”

Let us now see how this can be used to obtain an upper bound on the distance in the Schatten 1-norm, which is more operationally useful. Note that for a $d \times d$ matrix $X$,

$$||X||_2 \leq ||X||_1 \leq \sqrt{d}\ ||X||_2 \tag{19.13}$$

The latter inequality can be shown using the Cauchy-Schwarz inequality. Due to the factor of $\sqrt{d}$, to get a non-trivial bound, we often need $||X||_2$ to be exponentially small in the number of degrees of freedom.

Since the unitary operator does not act on $R$,

$$\sigma_R = \psi_R \tag{19.14}$$

As warm-up, let us find an upper bound on the 1-norm distance between $\sigma_2$ and the maximally mixed state $\tau_{A_2} = \frac{I_{A_2}}{d_{A_2}}$.

$$\begin{aligned} ||\sigma_{A_2} - \tau_{A_2}||_1^2 &\leq d_{A_2}\,\mathbb{E}\,||\sigma_{A_2} - \tau_{A_2}||^2 \\ &= d_{A_2}\,\mathbb{E}(\mathrm{tr}\sigma_{A_2}^2 - \frac{1}{d_{A_2}}) \\ &= d_{A_2}(\frac{p+q}{2} + \frac{p-q}{2}\mathrm{tr}\psi_A^2) - 1 \\ &= \frac{d_{A_2}}{d_{A_1}}\mathrm{tr}\psi_A^2 \\ &\leq \text{small if } d_{A_1} \gg d_{A_2} \end{aligned} \tag{19.15}$$

Similarly, the trace distance between $\sigma_{A_2R}$ and $\sigma_{A_2} \otimes \sigma_R$ is upper-bounded by

$$\begin{aligned} \mathbb{E}\,||\sigma_{A_2R} - \sigma_{A_2} \otimes \sigma_R||_1^2 &\leq d_{A_2}d_R\,\mathbb{E}\,||\sigma_{A_2R} - \sigma_{A_2} \otimes \sigma_R||_2^2 \\ &\leq \frac{d_{A_2}d_R}{d_{A_1}}(\mathrm{tr}\psi_{AR}^2 + \mathrm{tr}\psi_A^2\ \mathrm{tr}\psi_R^2) \end{aligned} \tag{19.16}$$

Let us now try to estimate the sizes of the different terms in (19.16).

Let us now take $n$ copies of our state. Take $|\psi\rangle$ to be a *typical purification* of $\rho_{AB}^{\otimes n}$. Given a purification $|\phi\rangle_{ABR}$ of $\rho_{AB}$, $|\psi\rangle$ is defined as

$$|\psi\rangle = c(\Pi^n_{\phi_A,\delta} \otimes \Pi^n_{\phi_B,\delta} \otimes \Pi^n_{\phi_B,\delta})\,|\phi\rangle_{ABR}^{\otimes n} \tag{19.17}$$

Then

$$\begin{aligned} \mathrm{tr}\psi_A^2 &\approx \exp(-nS(A)_\phi) = \exp(-nS(A)_\rho) \\ \mathrm{tr}\psi_R^2 &\approx \exp(-nS(R)_\phi) = \exp(-nS(AB)_\rho) \\ \mathrm{tr}\psi_{AR}^2 &\approx \exp(-nS(AR)_\phi) = \exp(-nS(B)_\rho) \geq \mathrm{tr}\psi_A^2\ \mathrm{tr}\psi_R^2 \end{aligned} \tag{19.18}$$

This means the second term in (19.16) can be ignored. Further, we can estimate

$$d_R \approx \exp(nS(AB)_\rho), \quad d_A \approx \exp(nS(A)_\rho) \tag{19.19}$$

So (19.16) is small if

$$\frac{d_A d_R}{d_{A_1^2}}\mathrm{tr}\psi_{AR}^2 \ll 1 \Rightarrow \log d_{A_1} \gg \frac{1}{2}\log(d_A d_R \mathrm{tr}\psi_{AR}^2) \approx \frac{1}{2}nI(A:R) \tag{19.20}$$

Let us now apply this to merging. The key idea is that due to the decoupling between $A_2$ and $R$ when $A_1$ consists of $\frac{1}{2}nI(A:R)$ qubits, the final state can be rotated to purify $R$ and $A_2$ separately.

We now have a "fully quantum Slepian wolf" protocol, where $\frac{1}{2}I(A:R)[q \to q]$ is used, to produce $\frac{1}{2}I(A:B)[qq]$.

Let us translate the bound obtained above for the trace distance to a bound on fidelity:

$$\frac{1}{2}||\sigma_{A_2R} - \sigma_{A_2} \otimes \sigma_R||_1 \leq \epsilon \Rightarrow F(\sigma_{A_2R}, \sigma_{A_2} \otimes \sigma_R) \geq 1 - \epsilon \tag{19.21}$$

We know that $\sigma_{A_2R}$ is purified by $U_{A\to A_1A_2}\left|\psi\right\rangle_{ABR}$, while $\sigma_{A_2} \otimes \sigma_R$ can be purified by $\left|\Phi\right\rangle_{A_2\tilde{B}} \otimes \left|\psi\right\rangle_{B_AB_BR}$. By Uhlmann's theorem, this implies that there exists a unitary acting on the complement of $A_2R$, $V_{A_1}B \to \tilde{B}B_AB_B$, that achieves this fidelity.

Implications of this discussion for relations between various protocols and resource inequalities in quant-ph/0606225:

Let us first try to express the Fully Quantum Slepian Wolf(FQSW)/ merging protocol as a resource inequality. Suppose we have an isometry $W$ from a source $S$ to $AB$. We denote with $\langle W_{S\to AB} : \psi_S\rangle$ the ability to produce the state $\rho_{AB}$ when the state on the source if $\psi_S$. Then we have

$$\langle W_{S\to AB} : \psi_S\rangle + \frac{1}{2}I(A:R)[q \to q] \geq \langle W_{S\to B_AB_B} : \psi_S\rangle + \frac{1}{2}I(A:B)[qq] \tag{19.22}$$

Using teleportation, we can equivalently write

$$\langle W_{S\to AB} : \psi_S\rangle + S(A|B)[q \to q] + I(A:B)[c \to c] \geq \langle W_{S\to B_AB_B} : \psi_S\rangle \tag{19.23}$$

Let us now run this protocol backwards. This gives us the FQRS, or the fully quantum reverse shannon protocol. While FQSW sends $\rho_{AB} \to \omega_{B'}$ where $B' = B_AB_B$, FQRS sends $\omega_{B'} \to \rho_{AB} = \mathcal{N}_{B'\to AB}(\omega)$, where $\mathcal{N}_{B'\to AB}$ is a channel from Bob ($B'$) to Alice($A$), where Alice keeps the environment. This can be seen as a protocol for *state splitting*,

$$\frac{1}{2}I(A:R)[q \leftarrow q] + \frac{1}{2}I(A:B)[qq] \geq \langle N_{B'\to AB} : \omega_B\rangle \tag{19.24}$$

Renaming the various parties, this can be written in a more standard form:

$$\frac{1}{2}I(A:B)[q \to q] + \frac{1}{2}I(B:E)[qq] \geq \langle N_{A'\to BE_A} : \omega_{A'}\rangle \tag{19.25}$$

Now recall the father protocol

$$\langle \mathcal{N}_{A'\to B} : \omega_{A'}\rangle + \frac{1}{2}I(A:E)[qq] \geq \frac{1}{2}I(A:B)[q \to q] \tag{19.26}$$

What if Alice keeps the environment? Then we have a channel $\mathcal{N}_{A'\to BE_A}$. Bob has purification, so $S(E)[qq]$ can be recovered. Net entanglement is

$$S(E) - \frac{1}{2}I(A:E) = \frac{1}{2}I(B:E) \tag{19.27}$$

$$\langle N_{A'\to BE_A} : \omega_{A'}\rangle = \frac{1}{2}I(A:B)[q \to q] + \frac{1}{2}I(B:E)[qq] \tag{19.28}$$

Shown in Devetak, quant-ph/0505138.

## 19.3 Merging and quantum error correction

Suppose we have a state with $a$ ebits between Alice and Rebecca, and $b$ ebits between Alice and Bob.

$$|\psi\rangle = |\phi\rangle_{A_1R}^{\otimes a} |\phi\rangle_{A_2B}^{\otimes b} \tag{19.1}$$

Merging requires $a$ qubits from $A$ to $B$, or $b$ qubits from $A \to R$. Both tasks are quite trivial, and can be accomplished by simply handing over the right qubits to Bob or Rebecca.

If instead we use a Haar-random unitary to accomplish this task, then we can use *any* $a + \delta$ qubits sent or $b + \delta$ qubits sent to $R$. But now, the *same* unitary works for both receivers, and it does not matter which qubits are sent to $B$ or $R$.

Conversely if Bob gets $a - \delta$ qubits, he is decoupled, or if Rebecca gets $b - \delta$ then she is decoupled, and the merging task cannot be accomplished.

Similar to a quantum error-correcting code.

Applications: Hayden and Preskill, "Black holes as mirrors" 0708.4025. Throw in one qubit. Can we recover it from the Hawking radiation?

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 20: November 10, 2020

*Lecturer: Aram Harrow* *Scribe: Yuan Lee, Shreya Vardhan*

In this lecture, we complete the discussion on merging and re-visit $k$-designs.

## 20.1 Black holes as mirrors

Black holes can be formed in general relativity from the gravitational collapse of a star in a pure state. Once the black hole is formed, it behaves in many respects like a thermal state: it emits radiation at a certain temperature, and can be associated with a thermodynamic entropy. However, the evolution from a pure state to a mixed state would violate the unitarity of quantum mechanics.

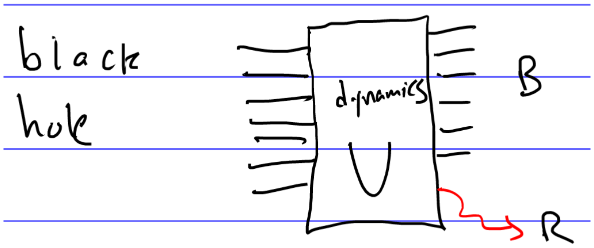
To see the same problem from another perspective, consider the following setup: Alice throws a diary into the black hole. Where does the information in the diary go? There are a few different options, but each of these presents some problems:

- The information is destroyed. This violates the unitarity of quantum mechanics.
- The information escapes gradually through the Hawking radiation. This violates the prediction from classical general relativity that nothing can escape from behind the horizon. Recently, a further problem with this option, known as the "firewall paradox", has also been discussed (see 1207.3123). This involves certain implications of monogamy of entanglement.
- The information escapes at the end. This violates the Bekenstein bound, which is a bound on the amount of entropy that can be contained within a given spatial region, as it implies that towards the end of the evaporation process, a very small region would have a very large entropy.
- There is a Planck-size black hole remnant left over at the end of the evaporation process. This violates the Bekenstein bound and destabilizes low-energy physics.
- A large black hole remnant is left over. This would imply that Hawking radiation stops being emitted at a relatively early time, which contradicts fairly reliable predictions of semiclassical gravity.

The difficulty of accepting any of these options lies at the heart of the conflict between general relativity and quantum mechanics.

It turns out (Hayden and Preskill, 0708.4025) that information discarded in an old black hole can be quickly recovered, given that we possess the Hawking radiation previously emitted from the black hole. This is a consequence of merging.

We can think of the black hole's time evolution as a unitary black box.



If the black hole starts off as a pure state, then the combined state $|\psi(t)\rangle_{BR}$ will be pure at all times $t$, if we assume the dynamics is unitary. Assuming that the time evolution operator $U(t)$ is Haar-random, the entropy of the black hole and the radiation are equal to $S(B)_{\psi(t)} = S(R)_{\psi(t)} = \min(|B|, |R|)$.

This gives rise to the Page curve (1301.4995): the entropy of the black hole increases until the Page time, after which its entropy decreases to zero.[2]

Now we consider the process of discarding information from Alice into an old black hole. To keep track of the information in Alice's qubits, we maintain a reference system that is (maximally) entangled to Alice's information.

Let the initial and final states be $\rho$ and $\sigma$ respectively. We now consider the state $\sigma_{NB'}$ that is shared by the black hole and the reference state. We expect this to be close to the maximally mixed state $\tau_N \otimes \tau_{B'}$. In fact, using the decoupling inequality from last lecture,

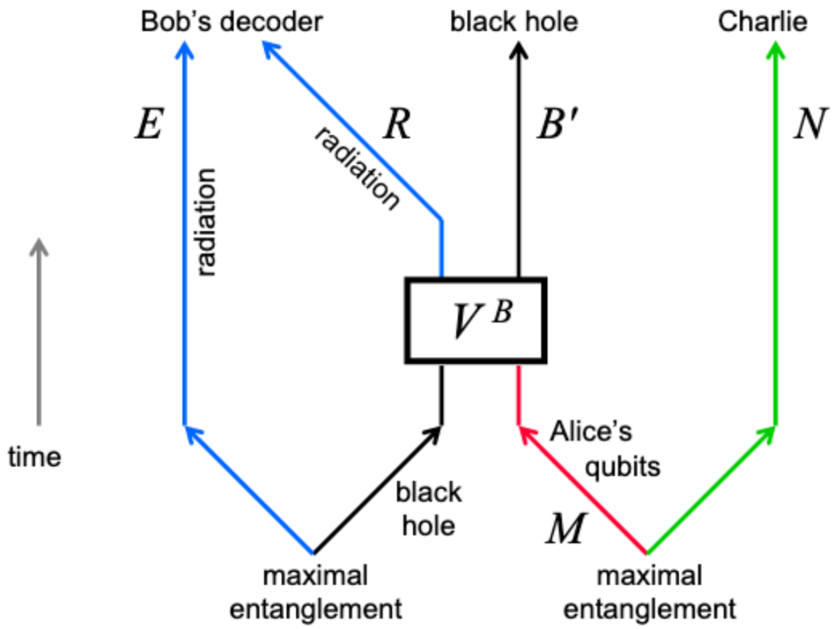
$$\mathbb{E}\|\sigma_{NB'} - \tau_N \otimes \tau_{B'}\|_1^2 \leq \frac{d_B d_B}{d_R^2} \mathrm{tr}\rho_{NB}^2 = \frac{d_N^2}{d_R^2}.$$

This distance is independent of the black hole dimension $d_B$.

In the above calculation, we use the fact that the black hole is old in the equation $\mathrm{tr}\rho_{NB}^2 = d_N/d_B$, because $\rho_{NB}$ would then be maximally-mixed.

Applying Uhlmann's theorem allows us to perform merging: we can reconstruct Alice's entanglement with the reference system using the radiation in $RE$.

[2] Aside: an article about black hole entropies was recently published in *Quanta*.



$\sigma$

$\rho$

## 20.2 Quantum capacity theorem (by merging)

Merging also gives a direct argument for the quantum capacity theorem (quant-ph/0702005).

As before, given a channel $N : A' \to B$, let its Stinespring dilation be $V_N : A' \to BE$. The coherent information is then $I_c(\phi_{A'}, N) = S(B)_{V(\phi)} - S(E)_{V(\phi)}$.

Let $|\gamma\rangle_{ABE} = (I_A \otimes V_{A' \to BE}) |\phi\rangle_{AA'}$. With $n$ copies of $|\gamma\rangle$, let $\Pi_{A^n \to S}$ be a projector onto a "typical subspace" $S$. In particular, we choose the projector $\Pi$ such that the projection onto the subsystem $S$ is maximally mixed. In other words, if $|\Psi\rangle_{SB^nE^n} = (\Pi_{A^n \to S} \otimes I_{B^nE^n}) |\gamma\rangle^{\otimes n}$ up to normalization, then $\Psi_S = I_S/d_S$. The state $\Psi$ would be close to the actual typical projection $\tilde{\Psi}$ in the subsytem $E$.

Then rank$\tilde{\Psi}_{E^n} \leq 2^{n(S(E)_\phi + \delta)}$ and tr$\tilde{\Psi}^2_{B^n} \leq 2^{-n(S(B)_\phi - \delta)}$.

Note that we can also write $|\Psi\rangle_{SB^nE^n} = (I_S \otimes V^{\otimes n}) |\Phi\rangle_{SS'}$, where $\Phi$ is the maximally-entangled state.

Like in Shannon's noisy coding theorem, we choose a random codespace $R$ using a fixed projector $P_{S \to R}$ and a Haar-random unitary $U_{S \to S}$, and conjugating $P_{S \to R}$ with $U_{S \to S}$. Then, the encoded state is

$$\begin{aligned} |\psi\rangle_{RB^nE^n} &= \sqrt{\frac{|S|}{|R|}} (PU \otimes I_{B^nE^n}) |\Psi\rangle_{SB^nE^n} \\ &= (I_R \otimes V^{\otimes n} U^\top) |\Phi\rangle_{RR'} . \end{aligned}$$

Here,

$$\mathbb{E}\|\tilde{\psi}_{RE^n} - \tilde{\psi}_R \otimes \tilde{\psi}_{E^n}\|_1^2 \leq d_R(\text{rank}\tilde{\psi}_{E^n})(\text{tr}\psi_{SE^n}^2) = d_R 2^{-n(S(B)_\phi - S(E)_\phi - 2\delta)}.$$

If $d_R \leq 2^{n(I_c(\phi,N)-3\delta)}$, the above distance is bounded above by $2^{-n\delta}$. Hence, for sufficiently small $\delta > 0$, the $R$ and $E^n$ subsystems are decoupled on average. We can strengthen this to worst-case decoupling by further restricting the codespace, allowing Bob to reconstruct Alice's state by merging.

## 20.3 Unitary $k$-designs

Let the matrix of all expected monomials $M(U) = U_{a_1b_1} \ldots U_{a_kb_k} U^*_{c_1d_1} \ldots U^*_{c_kd_k}$ under a distribution of unitaries $\mu$ be

$$\begin{aligned} G_\mu^k &= \mathbb{E}_{U\sim\mu}[(U \otimes U^*)^{\otimes k}] \\ &= \text{proj span } \{(I^{\otimes k} \otimes p_d(\pi)) \,|\Phi_d\rangle^{\otimes k} : \pi \in S_k\}, \end{aligned}$$

where $S_k$ is the set of permutations between $k$ qudits, $p_d(\pi)$ is the qudit permutation operator and $\Phi_d$ is the maximally-entangled state.

We say that a distribution $\mu$ on $U(d)$ is a *$k$-design* if

$$G_\mu^k = G_{\text{Haar}}^k.$$

Equivalently, for all matrices $\rho$,

$$\mathbb{E}_{U\sim\mu}[U^{\otimes k}\rho(U^\dagger)^{\otimes k}] = \mathbb{E}_{U\sim\text{Haar}} U^{\otimes k}\rho(U^\dagger)^{\otimes k}.$$

*1-designs* satisfy $\mathbb{E}[U\rho U^\dagger] = I/d$. We saw previously that the uniform distribution over the $d^2$ Pauli matrices form a 1-design. Moreover, the uniform distribution over a set of $O(d/\epsilon^2)$ random unitaries is an $\epsilon$-approximate 1-design. Drawing a random unitary or Pauli matrix is cheaper than drawing a random unitary from the $\epsilon$-net for $U(d)$, which has $(1/\epsilon)^{d^2}$ elements.

However, 1-designs are not enough for merging. For example, applying a random Pauli does not generate entanglement in an intially unentangled state, whereas applying a Haar-random unitary does.

It turns out that 2-designs are sufficient for most applications, including merging. An example of a 2-design is the uniform distribution over the set of Clifford operations.

To show that the Cliffords form a 2-design, first note that we can decompose all matrices into sums of Pauli matrices. Therefore, it suffices to consider the action of the Cliffords on the Paulis.

Next, let $C$ be a random Clifford on $n$ qubits, and let $p \in \{0,1,2,3\}^n$. Defining $\sigma_q = \sigma_{q_1} \otimes \ldots \otimes \sigma_{q_n}$,

$$C\sigma_p C^\dagger = \begin{cases} I & \text{if } p = 0^n, \\ \sigma_q & \text{for random } q \neq 0^n \text{ if } p \neq 0^n. \end{cases}$$

Therefore,

$$\mathbb{E}[(C\sigma_p C^\dagger) \otimes (C\sigma_q C^\dagger)] = \begin{cases} I & \text{if } p = q = 0^n, \\ \frac{1}{4^n-1} \sum_{r \neq 0^n} \sigma_r \otimes \sigma_r & \text{if } p = q \neq 0^n, \\ 0 & \text{if } p \neq q. \end{cases}$$

Note that $\sum_r \sigma_r \otimes \sigma_r = 2^n \text{SWAP}$ (where the sum includes $r = 0$). Therefore,

$$\mathbb{E}[(C\sigma_p C^\dagger) \otimes (C\sigma_q C^\dagger)] \in \text{span}\,\{I, \text{SWAP}\}.$$

This implies that the uniform distribution over the Clifford group is a 2-design, since SWAP commutes with $U \otimes U$.

In fact, the uniform distribution over the Cliffords is also a 3-design, but not generally a 4-design.

It is more expensive to draw a uniform sample from the Clifford group than the Pauli group, as the Clifford group has size $2^{n^2}$. However, it is still cheaper than drawing from an $\epsilon$-net of $U(d)$.

It turns out that we can generate approximate $k$-designs for any $k$ using a sufficiently large set of random unitaries. We find the number of unitaries needed using the matrix Chernoff bound, which states that

$$P\left( \|\frac{1}{n}\sum_{i=1}^{n} X_i\|_\infty \geq \delta \right) \leq 2de^{-n\delta^2},$$

where $X_1, \ldots, X_n$ are iid $d \times d$ matrices with mean zero and $\mathbb{E}\|X_i\|_\infty \leq 1$.

If $\mu$ is the uniform distribution over $m$ random unitaries $\{U_1, \ldots U_m\}$, it is convenient to define

$$X_i = (U_i \otimes U_i^*)^{\otimes k} - G^k_{\text{Haar}}.$$

Then $G^k_\mu - G^k_{\text{Haar}} = (1/m)\sum_{i=1}^m X_i$. Therefore,

$$\|G^k_\mu - G^k_{\text{Haar}}\|_\infty \leq \delta \text{ if } m = O(k(\log d)/\delta^2)$$
$$\Rightarrow \|G^k_\mu - G^k_{\text{Haar}}\|_1 \leq \delta \text{ if } m = O(d^{2k}k(\log d)/\delta^2).$$

We can get a lower bound on the rank $n$ of $\mathbb{E}[U^{\otimes k} |0\rangle \langle 0|^{\otimes k} (U^\dagger)^{\otimes k}] \approx \Pi_{\text{sym}}/\text{tr}\Pi_{\text{sym}}$ for any approximate $k$-design over $U$:

$$n \geq \text{tr}\Pi_{\text{sym}} = \binom{d+k-1}{k} = O(d^k).$$

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 21: November 12, 2020

*Lecturer: Aram Harrow* *Scribe: Shankar Balasubramanian*

## 21.1 Random states as quantum error correcting codes

Random states and QECC states are generically high entangled across cuts, and so one would expect random unitaries/states to serve as potentially good QECCs. Define the stabilizers $S = \langle S_1, S_2, \ldots, S_{n-k}\rangle$ and further define $W_\ell$ to be a string of Pauli matrices of total weight $\leq \ell$.

If $N(S) \cap W_\ell = \{I\}$ and $|\psi\rangle \in C$ where $C$ is the codespace, then $\psi_A$ is approximately maximally mixed for $|A| \leq \ell$. To this end, suppose $s_1, s_2, \ldots, s_n$ are random commuting Paulis. Then,

$$|W_\ell| \sim \sum_{k=1}^{\ell} \binom{n}{k} 3^k \approx \exp\left[n\left(H_2\left(\frac{\ell}{n}\right) + \frac{\ell}{n}\log 3\right)\right].$$

We have that $S = N(S)$, so

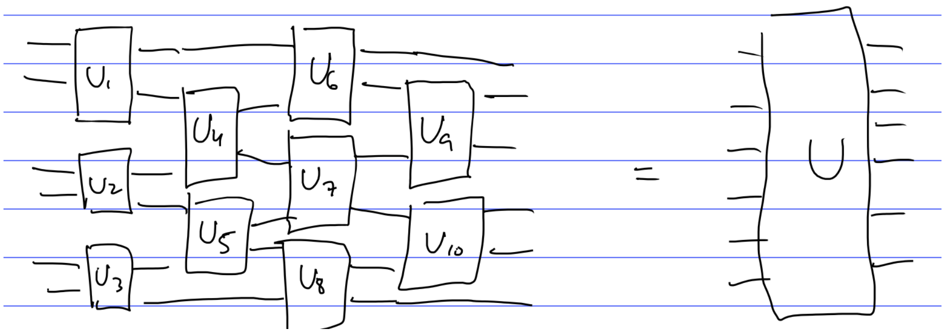
$$\mathbb{E}\,|S \cap W_\ell| = |W_\ell| 2^{-n} \ll 1$$

for $\ell/n < c$. This is known as a Gilbert-Varshamov bound. See quant-ph/0303022 for more details.

## 21.2 $k$-designs from random circuits

Let us consider the circuit construction shown below, where all of the unitaries $U_1$ through $U_{10}$ are chosen from a Haar random distribution over $U(d^2)$ or are a $k$-design on $U(d^2)$. What kinds of properties can such circuits have?

Let us restrict to the case where $k = 2$ and $d = 2$. We also define $p, q \in \{0, 1, 2, 3\}^2$.



Under the action of the quantum circuit, the local operator $\sigma_p \otimes \sigma_q$ becomes

$$\sigma_p \otimes \sigma_q \to \mathbb{E}_U (U \otimes U)(\sigma_p \otimes \sigma_q)(U \otimes U)^\dagger, \\ = \begin{cases} 0, & \text{for } p \neq q \\ I, & \text{for } p = q = 00 \\ \frac{1}{15}\sum_{r \neq 00} \sigma_r \otimes \sigma_r, & \text{for } p = q \neq 00 \end{cases}$$

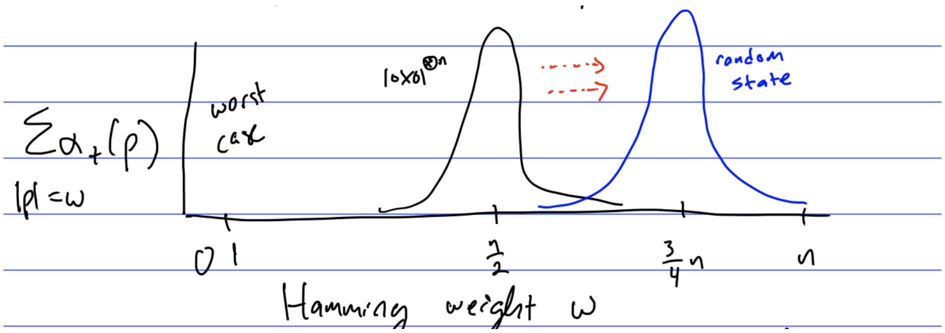
where we have used the fact that $M \to \mathbb{E}(U \otimes U)M(U \otimes U)^\dagger$ is a linear map. Let us start with an initial density matrix $\rho(0) = |0\rangle\langle 0|^{\otimes n}$; then the action of the circuit on this density matrix can be expressed by the relation $\rho(t) = U_t \rho(t-1) U_t^\dagger$ where $U_t$ are random 2-qubit gates on qubits $i$ and $j$. Because $U_t$ are 2-designs, we want to compute the quantity

$$\mathbb{E}[\rho(t) \otimes \rho(t)] = 2^{-n} \sum_p \alpha_t(p) \sigma_p \otimes \sigma_p + \sum_{p \neq q} \beta_t(p,q) \sigma_p \otimes \sigma_q. \tag{21.1}$$

For the second term, we know that the expectation value will render it zero (or vanishingly small) once all of the qubits have been touched a sufficient number of times. Therefore, this quantity can be modelled by the dynamics of $\alpha_t(p)$ by the relation $\alpha_t = M\alpha_{t-1}$ for $M$ a stochastic matrix. In particular, the stochastic update rule is to map $00 \to 00$ and $r \neq 00 \to r' \sim U[\{01, \ldots, 33\}]$ where $U[S]$ indicates a uniform distribution over elements in set $S$. Since $M$ has largest eigenvalue equal to 1, we can compute the steady state vector:

$$\alpha_t = \lim_{t \to \infty} M^t \alpha_0 = \begin{bmatrix} 2^{-n} \\ 4^{-n}/(1+2^n) \\ \vdots \\ 4^{-n}/(1+2^n) \end{bmatrix}.$$

Now we ask what the mixing time of this Markov chain is. The worst starting point is the string $1000\ldots0$, because the only transition can occur when qubits 0 and 1 are touched, which only locally updates the string. Therefore, it takes a long time to reach



all bits. In general, the convergence time depends on the geometry. For 1D circuits, $O(n^2)$ gates are required. For a fully connected architecture, the convergence time is $n + n/2 + n/3 + \ldots \sim n \log n$ gates are required.

Consider the state $|0\rangle\langle 0|^{\otimes n} = \left(\frac{I+\sigma_3}{2}\right)^{\otimes n}$. This can be thought of as a mixture of $\{00\ldots 00, 00\ldots 03, 00\ldots 30, \cdots, 33\ldots 33\}$. The time requires for the circuit to be maximally entangled from this starting state is $\Omega(n^2)$ for 1D circuits and $\Omega(n \log n)$ for fully connected circuits. Thus, even for this simplified case, we get the same bounds. We can plot the distribution of outcomes, shown above.

Let us explore the structure of the Markov chain more. Denote $* = \{1, 2, 3\}$ which collapses the state space to 00, $0*$, $*0$, and $**$. The transition probabilities (written above the arrows) are

$$p_i p_j(t) = \{00\} \xrightarrow{1} p_i p_j(t+1) = \{00\}$$
$$p_i p_j(t) = \{0*, *0, **\} \xrightarrow{\frac{6}{15}} p_i p_j(t+1) = \{0*, *0\}$$
$$p_i p_j(t) = \{0*, *0, **\} \xrightarrow{\frac{9}{15}} p_i p_j(t+1) = \{**\}$$

Therefore, the Hamming weight $w$ follows a random walk with a drift towards $w = \frac{3}{4}n$. This also goes under the name Ornstein-Uhlenbeck process. It can be shown that

$$\left| w - \frac{3}{4}n \right| \sim e^{-ct/n}$$

for $w > 0.1n$. The random walk is mixed when $\left|w - \frac{3}{4}n\right| \leq \sqrt{n}$ or when $t = O(n \log n)$.

It is important to beware that the convergence time depends on the metric used. For example, define the anticoncentration

$$\Lambda = \sum_{z \in \{0,1\}^n} |\langle z|U|0^n\rangle|^4 = \sum_z p(z)^2.$$

It can be computed that $\Lambda = 1$ for $U = I$ and that $\Lambda = 2^{-n}$ for Haar random $U$. The computation goes like

$$\begin{aligned}\mathbb{E}\Lambda &= \mathrm{Tr}\left(\sum_z |z\rangle\langle z| \otimes |z\rangle\langle z|\right) \frac{I+F}{2^n(2^n+1)} \\ &= \frac{2}{2^n+1}.\end{aligned}$$

Anticoncentration is a property that appears in the goal of quantum supremacy. In general, the hardness of simulating quantum circuit families uses anticoncentration to extend worst-case hardness to average-case hardness.

Alternatively, we may also write

$$\sum_z |z\rangle\langle z| \otimes |z\rangle\langle z| = 2^{-n} \sum_{p\in\{0,3\}^n} \sigma_p \otimes \sigma_p$$

and this imples

$$\Lambda = \sum_{p\in\{0,3\}^n} \alpha_t(p).$$

In 1D, $\Lambda$ converges in $O(n \log n)$ steps, or in $O(\log n)$ depth. The details can be found in (2005.02421)

## 21.3 Techniques for bounding convergence time

The first method is the spectral method, where one computes the value of the second largest eigenvalue, so that the error as a function of iteration $t$ vanishes as $\sim (\lambda_2(M))^t$. For $k > 2$, $M$ is no longer stochastic. Let us restrict to 1D and compute

$$G^k_{\text{circuit}} = \mathbb{E}_{1\le i\le n-1}\mathbb{E}_{U\sim U(4)}\left(\left(I^{\otimes i-1}\otimes U \otimes I^{\otimes n-i-1}\right)\otimes\left(I^{\otimes i-1}\otimes U^* \otimes I^{\otimes n-i-1}\right)\right)^{\otimes k}.$$

Then, under many iterations

$$(G^k_{\text{circuit}})^t \to G^k_{\text{Haar}} = \text{proj}\{|\psi\rangle : \left(U^{\otimes k}\otimes U^{*\otimes k}\right)|\psi\rangle = |\psi\rangle\}$$

and we find

$$G^k_{\text{ciruit}} = \frac{1}{n-1}\sum_{i=1}^{n-1} P_{i,i+1},$$

which has eigenvalues in the interval $[0,1]$. We may write $G^k_{\text{ciruit}}$ in the block diagonal structure

$$G^k_{\text{ciruit}} = \left[\begin{array}{c|c} G^k_{\text{Haar}} & \\ \hline & A \end{array}\right]$$

with $\| A \|_\infty < 1$. There are several bounds and conjectures for the maximum singular value of $A$. Currently, in (1208.0692), it is rigorously shown that

$$\| A \|_\infty \leq 1 - \frac{1}{nk^{O(1)}},$$

while (1905.12053) argues that it is possible for the upper bound to be independent of $k$.

Another method that exists is a method of mapping quantum circuits to statistical mechanical models, some of which possess exact solutions. We first start by computing the expectation value of the quantity

$$\mathbb{E}[(U \otimes U^*)^{\otimes k}],$$

where $U$ is Haar random. It can be verified that

$$\begin{aligned}\mathbb{E}[(U \otimes U^*)^{\otimes k}] &= \text{proj span}\left\{|\Phi_\pi\rangle = (I^{\otimes k} \otimes P_d(\pi))|\Phi\rangle^{\otimes k}\right\} \\ &= \sum_{\sigma,\tau} |\Phi_\sigma\rangle\langle\Phi_\tau| \text{Wg}(\sigma, \tau).\end{aligned}$$

In general, what is proj span $\{|v_1\rangle, |v_2\rangle, \ldots, |v_m\rangle\} = \Pi$? Define

$$K = \sum_{i=1}^{m} |v_i\rangle\langle i|.$$

Then, it follows that

$$\Pi = K(K^\dagger K)^{-1} K^\dagger.$$

This can be seen since $\Pi|v_i\rangle = \Pi K|i\rangle = K|i\rangle = |v_i\rangle$, as desired. In the current case,

$$K = \sum_{\pi} |\Phi_\pi\rangle\langle\pi|,$$

and

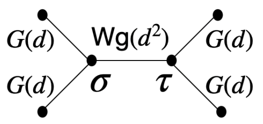
$$K^\dagger K = \sum_{\sigma,\tau} \langle\Phi_\sigma|\Phi_\tau\rangle |\tau\rangle\langle\sigma|.$$

Let us carefully compute

$$\begin{aligned}G_{\sigma,\tau} = \langle\Phi_\sigma|\Phi_\tau\rangle &= \langle\Phi|^{\otimes n} I^{\otimes n} \otimes P_d(\sigma^{-1}\tau)|\Phi\rangle^{\otimes n} \\ &= \frac{\text{Tr} P_d(\sigma^{-1}\tau)}{d^n} = d^{\#\text{ cycles}(\sigma^{-1}\tau) - n} = d^{-\text{dist}(\sigma,\tau)}.\end{aligned}$$

We define the Weingarten function $\text{Wg} = G^{-1}$ and it is close to the identity if $d \gg n^2$. Schematically, we may express the relation between the original expectation value and the Weingarten function by the diagram below (see 1905.12053 for more details):

$$\mathbb{E}[(U \otimes U^*)^{\otimes k}] = \sum_{\sigma,\tau}$$



**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 22: November 17, 2020

*Lecturer: Aram Harrow* *Scribe: Shankar Balasubramanian*

## 22.1 Clarifications in previous proof

An important identity that we made use of is

$$\mathrm{Tr}(\text{cycle}) = \sum_{i_1.i_2,\cdots,i_n} \mathrm{Tr}\left(|i_1, i_2, \cdots, i_n\rangle\langle i_2, i_3, \cdots, i_n, i_1|\right) = d$$

The notation $\mathrm{dist}(\sigma, \tau)$ means the number of transpositions required to get from $\sigma$ to $\tau$. The identity $\mathrm{dist}(\sigma, \tau) = n - \#\ \text{cycles}(\sigma^{-1}\tau)$ follows from this definition.

Finally, as a remark, the matrix $K^\dagger K$ is often called the Gram matrix.

## 22.2 $n = 2$ case

Let us construct the Gram matrix when $n = 2$ which is $G_{\sigma,\tau} = \langle \Phi_\sigma | \Phi_\tau \rangle$. This matrix looks like

$$G = \begin{bmatrix} 1 & 1/d \\ 1/d & 1 \end{bmatrix} = \left(1 + \frac{1}{d}\right)|+\rangle\langle+| + \left(1 - \frac{1}{d}\right)|-\rangle\langle-|.$$

The Weingarten function is the inverse of this matrix, or

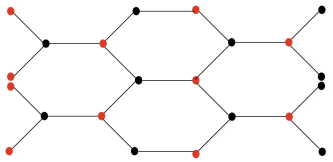
$$\mathrm{Wg} = \frac{d^2}{d^2-1}\begin{bmatrix} 1 & -1/d \\ -1/d & 1 \end{bmatrix} = \frac{d}{d+1}|+\rangle\langle+| + \frac{d}{d-1}|-\rangle\langle-|.$$

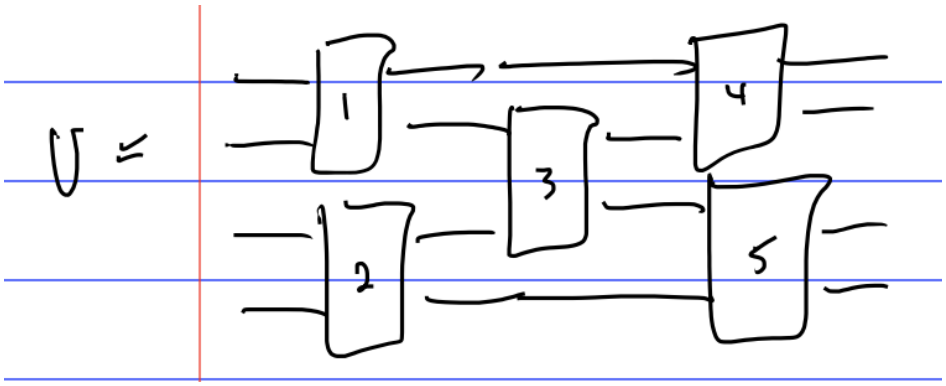
Because the off-diagonal elements are negative, we encounter a sign problem, which we need to resolve in order to write a random circuit in terms of the partition function of a statistical mechanical model.

To see how to correct the sign problem, consider the example circuit shown in the handwritten figure below

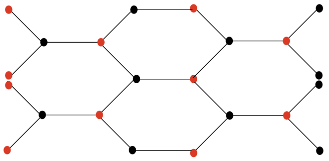
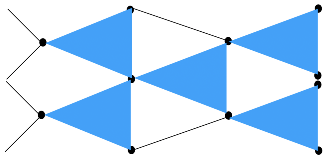
We may write the expression

$$\mathbb{E}(U \otimes U^*)^{\otimes 2} = \sum_{\bullet \in \{I,F\}}$$

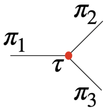


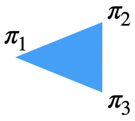


using the diagrammatic approach introduced in the last lecture. Now, we perform a step called decimation, whereby we sum over the red dots only. Remarkably, this allows us to get rid of the sign problem.

$$\sum_{\bullet,\bullet\in\{I,F\}}$$

$$= \sum_{\bullet\in\{I,F\}}$$

.

In this new notation, we have defined

$$\sum_{\tau}$$


$$= \sum_{\tau} \mathrm{Wg}_{\tau,\pi_1}(d^2) G_{\tau,\pi_2}(d) G_{\tau,\pi_3}(d) =$$


.

We can split the evaluation up into several cases, corresponding to whether the permutations $\pi_i$ are $I$ or $F$. Let us denote the weight of the triangle diagram $D_{\vec{\pi}}$, where $\vec{\pi}$ is a vector of the three permutations. Then, we have:

$$(\pi_1,\pi_2,\pi_3) \in \{(I,I,I),(F,F,F)\} \to D_{\vec{\pi}} = \frac{d^4}{d^4-1}\left(1 - \frac{1}{d^2}\cdot\frac{1}{d}\cdot\frac{1}{d}\right) = 1,$$

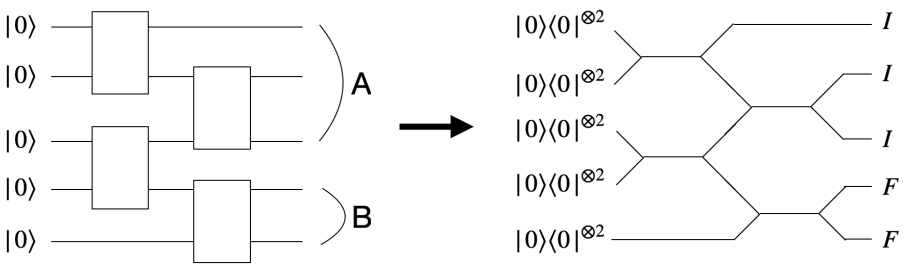
as well as

$$(\pi_1,\pi_2,\pi_3) \in \{(I,F,F),(F,I,I)\} \to D_{\vec{\pi}} = \frac{d^4}{d^4-1}\left(\frac{1}{d^2} - \frac{1}{d}\cdot\frac{1}{d}\right) = 0.$$

and

$$(\pi_1,\pi_2,\pi_3) \in \{(I,I,F),(I,F,I),(F,F,I),(F,I,F)\} \to D_{\vec{\pi}} = \frac{d^4}{d^4-1}\left(\frac{1}{d} - \frac{1}{d^3}\cdot\frac{1}{d}\right) \sim \frac{1}{d}.$$

Notice that all of these weights are positive. Therefore, it suffices to write the random circuit amplitude as a sum over all possible assignments of permutations on the black dots with the weight of a particular assignment being the product of the corresponding weights of all of the triangles. This can be written as the partition function of some effective spin model.



## 22.3 Estimating entanglement

As an application, we can use these mappings as a tool for understanding entanglement across a cut in a random unitary circuit. Consider the following circuit above and its dual mapping; the regions $A$ and $B$ are subsystems of the output density matrix.

We are interested in computing the quantity

$$Z = \mathbb{E}[\mathrm{Tr}(\rho_B^2)],$$

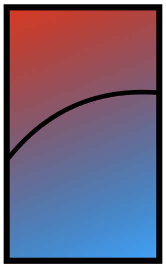
which is exactly equal to the types of expectation values we were considering. To understand how to calculate this using the statistical mechanical model formalism, we first note that in the triangle diagrams, if the two of the $\pi$'s on the right side of the triangle are the same, then the $\pi$ on the left corner must be equal to the $\pi$'s on the right. If the two $\pi$'s disagree, then the $\pi$ on the left can take either of the two values with weight $\sim 1/d$. The latter case corresponds to the weight of creating a single domain wall within the triangle (here, a domain wail refers to a cut for which permutations are given different assignments on either side of the cut).

Thus, since the nodes on the right end are forced to be $I$ in region $A$ and $F$ in region $B$, the dominant assignment of permutations to the internal nodes (i.e. the saddle point value of the partition function) occurs when one draws a domain wall that cuts the circuit into a red and blue region.

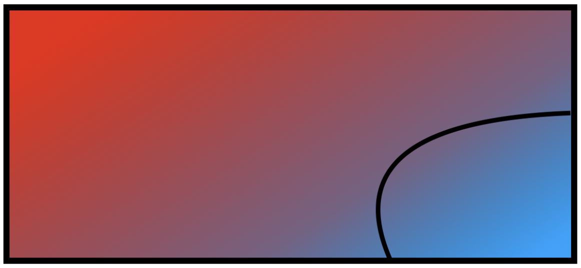
There are two limits of interest. Call the depth of the circuit $t$. When $t < n$, then the domain wall does not form completely and so the dominant configuration corresponds to the domain wall cutting across from the left to the right side of the circuit. The weight of this configuration scales like $(1/d)^t$, so

$$Z = \mathbb{E}[\mathrm{Tr}(\rho_A^2)] = \mathbb{E}\left[2^{-S_2(\rho_A)}\right] \approx \exp\left(t \log \frac{d^2+1}{2d}\right),$$

which is sensible because the circuit has not fully entangled the initial density matrix, so by increasing the depth by one unit, the entropy should increase by roughly $\log d$. When $t > n$, then the circuit is fully scrambled and the domain wall will stretch



$t < n$



$t > n$

from the upper/lower ends of the circuit to the right side. In this case $Z \approx d^{-n/2}$ or $S_2 = \frac{n}{2} \log d$ which implies that the entanglement has saturated. These cases are visually depicted in the figure above.

See section IV.A of 1804.09737 (originally 1608.06950) for more information.

Recent work in this area includes the following:

- Unitary + measurement circuits: Here, one alternates between applying columns of random unitaries in a brickwork architecture with random measurements inserted between adjacent columns. Here, a phase transition from a volume law (maximally entangled) phase to an area law (product state) phase is found as a function of the rate of measurement

- Random tensors and random unitary circuits are candidates for establishing quantum supremacy

- Computing $S_k$ for $k \geq 2$ and analytic continuation to $k = 1$ using the replica trick

## 22.4 Monogamy of entanglement

A simple definition of monogamy of entanglement is that if Alice and Bob are maximally correlated, then neither of them can be correlated with a third party. We will better try to understand this concept through two methods:

- Using symmetry (de Finetti theorems)

- Using information theory (approximate Markov states)

A motivating example is mean field theory. We start out with the Hamiltonian

$$\mathcal{H} = \sum_{i \sim j} h_{ij}$$

where $h$ are local Hamiltonians and $i \sim j$ means that $i$ and $j$ are neighbors. We will next make a mean field approximation where we assume that the Hamiltonian is close enough to one in which each particle interacts evenly among all other neighbors:

$$\mathcal{H} \approx \mathcal{H}_{MF} = \frac{D}{n} \sum_{1 \leq i < j \leq n} h_{ij}$$

where $D$ is the number of neighbors. For example, if $h_{ij} = F_{ij}$, then the ground state is the singlet state $\frac{|01\rangle - |10\rangle}{\sqrt{2}}$. If $n = 3$, we call the system frustrated because we cannot have all pairs of particles forming singlets with each other.

We claim that the ground state of $\mathcal{H}_{MF}$ look like $\rho^{\otimes n}$.

To see this, we note that $[\mathcal{H}_{MF}, P_d(\pi)] = 0$ for all $\pi \in S_n$, since the mean field Hamiltonian is invariant under swapping the particles. Therefore, we may write

$$\mathrm{Tr}(\mathcal{H}_{MF}\psi_{gs}) = \mathrm{Tr}\left(\mathcal{H}_{MF} \frac{1}{n!} \sum_{\pi \in S_n} P_d(\pi)\psi_{gs}P_d(\pi)^\dagger\right).$$

Now define

$$\omega = \frac{1}{n!} \sum_{\pi \in S_n} P_d(\pi)\psi_{gs}P_d(\pi)^\dagger.$$

Thus, $[\omega, P_d(\pi)] = 0$ and thus we can choose the ground state to be symmetric WLOG. This does not prove that it must be a tensor power state, and we will finish the proof next time.

Why not use the pure state

$$|\psi\rangle \propto \sum_{\pi \in S_n} P_d(\pi)|\psi_{gs}\rangle,$$

which is also a valid eigenstate that is symmetric? The problem is that this can be zero, so it is safer to use density matrices over quantum states.

The de Finetti theorem states that

$$\omega_{ij} = \int d\mu(\rho)\, \rho^{\otimes 2}.$$

Why do we need a mixture (the integral)? An example would be the cat state $(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})/\sqrt{2}$, which is not close to a product state and can only be expressed as a mixture.

The original classical version of de Finetti's theorem was proved in 1931 by Bruno de Finetti. It states that suppose $p$ is an infinitely exchangeable probability distribution, or

$$p(x_1, x_2, \cdots) = p(x_{\pi(1)}, x_{\pi(2)}, \cdots)\, \forall \pi.$$

Then $\exists \mu$ such that $\forall k$

$$p(x_1, x_2, \cdots, x_k) = \int d\mu(q)\, q(x_1) \cdots q(x_k),$$

or equivalently $p_{1\ldots k} = \int d\mu(q)\, q^{\otimes k}$. So naively, permutation invariance is not the same as independence, because we can allow for mixtures of IID distributions and still preserve permutation invariance. In 1980, Diaconis and Freedman made this result more quantitative. In particular, they found that if

$$p(x_1, \cdots, x_n) = p(x_{\pi(1)}, \cdots, x_{\pi(n)})\, \forall n,$$

then

$$\frac{1}{2} \| \, p_{1\ldots k} - \int d\mu(q)\, q^{\otimes k} \, \|_1 \leq \min\left( \frac{k(k-1)}{2n}, \frac{k|x|}{n} \right).$$

In 2002, Caves Fuchs, and Schack derived a quantum version of de Finetti's theorem. We will follow the treatment in Chiribella (2010), 1010.1875.

We first purify $\rho_{A_1, \ldots, A_n} \to |\psi\rangle_{A_1, B_1, \ldots, A_n, B_n} \in \mathrm{Sym}^n\, \mathbb{C}^{d^2}$. Thus, it suffices to prove the de Finetti theorem for pure states.

**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

Lecture 23: Nov 19, 2020

*Lecturer: Aram Harrow* *Scribe: Matthew Kearney, Andrew Tan*

# 23.1 de Finetti Thm for Pure Symmetric States

Let $|\psi\rangle$ be a state in the symmetric subspace of $n+k$ systems in $d$ dimensions.

$$|\psi\rangle \in \mathrm{Sym}^{n+k}\mathbb{C}^d$$

We will show that

$$F(\mathrm{tr}_n(\psi), \int d\mu(\phi)\phi^{\otimes k})^{2k} \geq 1 - \frac{kd}{n}$$

## 23.1.1 Tomography

We will first take a detour and talk briefly about tomography, which describes which measurements to make to estimate a state. Specifically we can ask the question, given $|\phi\rangle^{\otimes n}$ how well can we estimate $|\phi\rangle$? We will measure with a continuously indexed set of POVM: $\{M_{\hat{\phi}}\}_{\hat{\phi}}$ such that the following holds:

$$\int d\hat{\phi} M_{\hat{\phi}} = \Pi_{\mathrm{sym}}$$

Unlike the usual case where we need the measurements to sum to the identity, since $|\phi\rangle^{\otimes n}$ lies in the symmetric subspace, we only need the measurement operators to sum to the projector onto this subspace. It turns out that the optimal set of measurements is given by

$$M_{\hat{\phi}} = c\hat{\phi}^{\otimes n}$$

Where we can calculate the constant c by calculating

$$\int d\hat{\phi}\hat{\phi}^{\otimes n} = \frac{\Pi_{\mathrm{sym}}}{\mathrm{tr}(\Pi_{\mathrm{sym}})} = \frac{\Pi_{\mathrm{sym}}}{\binom{d+n-1}{n}}$$

This gives us

$$c = \binom{d+n-1}{n}$$

We can now return to the proof of the de Finetti Theorem. Recall that the squared fidelity for pure states is given by

$$F(\phi, \hat{\phi}) = \mathrm{tr}\phi\hat{\phi}$$

Therefore we can calculate the expectation of the squared fidelity as

$$\begin{aligned}
\mathbb{E}F(\phi,\hat{\phi})^2 &= \int d\hat{\phi}\,\mathrm{tr}(\phi\hat{\phi})\,\mathrm{tr}(M_{\hat{\phi}}\phi^{\otimes n}) \\
&= \int d\hat{\phi}\,\mathrm{tr}(\phi\hat{\phi})\binom{d+n-1}{n}\mathrm{tr}(\phi\hat{\phi})^n \\
&= \binom{d+n-1}{n}\int d\hat{\phi}\,\mathrm{tr}(\phi\hat{\phi})^{n+1} \\
&= \binom{d+n-1}{n}\mathrm{tr}(\phi^{\otimes n+1}\int d\hat{\phi}\,\hat{\phi}^{\otimes n+1}) \\
&= \binom{d+n-1}{n}\mathrm{tr}(\phi^{\otimes n+1}\frac{\Pi_{\text{sym}}}{\binom{d+n}{n+1}}) \\
&= \frac{\binom{d+n-1}{n}}{\binom{d+n}{n+1}}\mathrm{tr}(\phi^{\otimes n+1}\Pi_{\text{sym}}) \\
&= \frac{\binom{d+n-1}{n}}{\binom{d+n}{n+1}}\mathrm{tr}(\phi^{\otimes n+1}) \\
&= \frac{\binom{d+n-1}{n}}{\binom{d+n}{n+1}} \\
&= \frac{n+1}{n+d} \\
&\geq 1-\frac{d}{n}
\end{aligned}$$

We can also calculate higher moments of the fidelity as well, in which case we find

$$\mathbb{E}F(\phi,\hat{\phi})^{2k} = \frac{\binom{d+n-1}{n}}{\binom{d+n-1+k}{n+k}} = \frac{(n+1)\ldots(n+k)}{(n+d)\ldots(n+d+k-1)} \geq 1-\frac{dk}{n}$$

### 23.1.2 de Finetti Theorem Proof

Let $|\psi\rangle$ be given by

$$|\psi\rangle = \int a_\phi \, |\phi\rangle^{\otimes n+k} \; d\phi$$

If we then trace over n of the qudits we get

$$\mathrm{tr}_n(\psi) = \int d\phi (M_\phi \otimes I^{\otimes k})\psi = \int d\phi \, p_\phi \psi_\phi$$

We would like to claim that $\psi_\phi \approx \phi^{\otimes k}$.

Calculating the fidelity we get

$$\begin{aligned}
F(\mathrm{tr}_n(\psi), \int d\phi \, p_\phi \phi^{\otimes k})^2 &\geq \int d\phi \, p_\phi F(\psi_\phi, \phi^{\otimes k})^2 \\
&= \int d\phi \, \mathrm{tr}(p_\phi \psi_\phi \phi^{\otimes k}) \\
&= \int d\phi \, \mathrm{tr}((M_\phi \otimes \phi^k)\psi) \\
&= \mathrm{tr}(\int d\phi \binom{d+n-1}{n} \phi^{\otimes n+k} \psi) \\
&= \frac{\binom{d+n-1}{n}}{\binom{d+n-1+k}{n+k}} \\
&= \frac{(n+1)\dots(n+k)}{(n+d)\dots(n+d+k-1)} \\
&\geq 1 - \frac{dk}{n}
\end{aligned}$$

As a corollary to this we have that for non pure states: If $\rho \in D_{d^{n+k}}$ and $[\rho, p_d(\pi)] = 0 \forall \pi$ then

$$F(\mathrm{tr}_n(\rho), \int d\mu(\sigma) \, \sigma^{\otimes k})^2 \geq 1 - \frac{d^2 k}{n}$$

### 23.1.3 Examples of de Finetti Theorem

Below are some examples to illustrate the theorem.

**Example 1**

Let $|\psi\rangle = (|0\rangle^{\otimes n} + |1\rangle^{\otimes n})/\sqrt{2}$. Then we have that for $k \geq 1$

$$\mathrm{tr}_{n-k}\psi = \frac{|0\rangle\langle 0|^{\otimes k} + |1\rangle\langle 1|^{\otimes k}}{2}$$

So we can start with a state that looks nothing like a product state and after just tracing out one system we end with a state that is exactly the state described by the de Finetti theorem.

**Example 2**

Let $\rho = \binom{n}{n/2}^{-1} \sum\limits_{x \in \{0,1\}^n, |x| = n/2} |x\rangle\langle x| := W^n_{n/2}$, which is far from any $\sigma^{\otimes n}$. Then we have that

$$\mathrm{tr}_{n-k}\rho = \sum_{j=0}^{k} \frac{\binom{n/2}{j}\binom{n/2}{k-j}}{\binom{n}{n/2}} W^k_j \approx 2^{-k}\binom{k}{j} W^k_j$$

**Example 3**

This example shows why it is important that $n$ must be bigger than $d$. We have that $|\phi\rangle = n \sum\limits_{\pi \in S_n} \mathrm{sgn}(\pi)\, |\pi(1), \ldots, \pi(n)\rangle \in \mathbb{C}^{\ltimes \otimes n}$.

$$\mathrm{tr}_{n-2}\phi = \binom{n}{2}\Pi_{\mathrm{anti}}$$

This state is very far from separable states even though $k = 2$.

## 23.2 Applications

### 23.2.1 Mean-Field Theory

Let the Hamiltonian acting on our set of qudits be given by

$$H = \binom{n}{k} \sum_{i_1 < i_2 < \cdots < i_k} h_{i_1, \ldots, i_k}$$

If we then look at the trace of the Hamiltonian applied to the ground state we get

$$\begin{aligned}\operatorname{tr}(H\psi_{gs}) &= \operatorname{tr}(H\rho)\\ &= \operatorname{tr}(h\otimes I^{\otimes n-k})\rho\\ &\geq \operatorname{tr}(h\int d\mu(\sigma)\sigma^{\otimes k}) - \|h\|_\infty \frac{d^2k}{n-k}\\ &\geq \min_\sigma \operatorname{tr}(h\sigma^{\otimes k}) - \|h\|_\infty \frac{d^2k}{n-k}\end{aligned}$$

At the same time we have that because $\psi_{gs}$ is the ground state we know

$$\operatorname{tr}(H\psi_{gs}) \leq \min_\sigma \operatorname{tr}(h\sigma^{\otimes k}) = \min_\sigma \operatorname{tr}(H\sigma^{\otimes n})$$

### 23.2.2 Security of QKD

In QKD, Alice sends $H^{a_1}|r_1\rangle \otimes H^{a_2}|r_2\rangle \otimes \cdots \otimes H^{a_n}|r_n\rangle$ for $a, r \in \{0,1\}^n$ uniformly random binary strings. Bob then applies $H^{b_1}\otimes H^{b_2}\otimes\cdots\otimes H^{b_n}$ for $b \in \{0,1\}^n$ also a random binary string. Against i.i.d. attacks (attacks where Eve does the same thing to every qubit), this protocol can tolerate a bit error rate $< p_c \approx 0.14$.

What to do about general attacks though? Alice and Bob can use a symmetric protocol therefore discarding $n-k$ quibits and leaving the remaining qubits in a state approximately equal to $\int d\mu(\sigma)\,\sigma^{\otimes k}$ with error $k/n$. In other words we can sacrifice $O(1/\epsilon^2)$ qubits to learn $\sigma$ to error $\epsilon$. Normally in cryptography we expect that security should be exponentially good in the amount of effort made, but here we can only keep $O(\sqrt{n})$ qubits and the error decreases as O(1/n).

### 23.2.3 Exponential de Finetti Theorem

Sometimes other theorems help us to get better bounds. The exponential de Finetti theorem states that if $\rho_{n+k} \in D_{d^{n+k}}$ symmetric, then

$$\rho_k = \operatorname{tr}_n(\rho_{n+k}) \approx \int d\mu(\sigma)\,\sigma^{k-r}\otimes\phi_r$$

Where $\phi_r$ is just an arbitrary density matrix on $r$ systems. In this case we find that the error is approximately less than or equal to $k^{O(d)}\exp\frac{-kr}{n+k}$.

### 23.2.4 de Finetti Reductions

If $\rho_n \in D_{d^n} \in \mathrm{Sym}$

$$\rho_k \leq (n+1)^{d^2} \int\limits_{\sigma \in D_d} \sigma^{\otimes n}\, d\sigma$$

Then we get that for some bad event B,

$$\mathbb{P}(B) = \mathrm{tr}(M\rho_n) \leq (n+1)^{O(d^2)} \int d\sigma\, \mathrm{tr}(M\sigma^{\otimes n})$$

If our probability is exponentially small, paying a polynomial pre-factor won't matter, so this can be useful for upper bounds.

### 23.2.5 Applications to Classical Optimization Algorithms

The goal of the optimization algorithms is to find

$$h_s(y) = \max_{x \in S} \langle x, y \rangle$$

For density matrices $D_d$ and measurement $M$ we have

$$h_{D_d}(M) = \|M\|_\infty$$

Solving the following is much harder however

$$h_{\mathrm{sep}}(M) = \max_{\alpha, \beta \in D_d} \mathrm{tr}(M(\alpha \otimes \beta))$$

It is NP hard to get error O(1/d). Define the following set

$$\mathrm{SepSym}(d,k) = \mathrm{conv}\{\sigma^{\otimes k} : \sigma \in D_d\}$$

We have that

$$\mathrm{SepSym}(d,k) \subseteq \mathrm{SymExt}(d,k,n) = \{\rho_k \in D_{d^k} : \exists \rho_n \in D_{d^n},\ \text{symmetric}\}$$

We have that

$$h_{\text{SepSym}}(M) \le h_{\text{SymExt}} \le h_{\text{SepSym}}(M) + O(\frac{d^2 k}{n})$$

### 23.2.6 Monogamy Using Information Theory

Let system A be entangled with systems $B_1, \dots, B_n$ such that $B_i$ is conditionally independent of $B_j$ given $A$ for $i \neq j$. There is a trade off between entanglement of $\rho_{AB_1}, \cdots, \rho_{AB_n}$ without requiring symmetry assumptions and $n \approx \log d$ instead of poly$(d)$.

We have that

$$\begin{aligned} 2 \log d_A &\ge I(A : B_1, \dots, B_n) \\ &= I(A : B_1) + I(A : B_2 | B_1) + \cdots + I(A : B_n | B_1 \dots B_{n-1} \end{aligned}$$

$$\mathbb{E}_{j \in [n]} I(A : B_j | B_1^{j-1}) = 2 \log(d_A)/n \le \epsilon^2 \quad \text{if } n \ge \frac{2 \log(d_A)}{\epsilon^2}$$

Why is this helpful? squashed entanglement:

$$E_{sq}(\rho^{AB}) - \inf\{\frac{1}{2} I(A : B|E) : \rho^{ABE} \text{ an extension of } \rho^{AB}\}$$

Then we have that $\mathbb{E}_i E_{sq}(\rho^{AB_i}) \le \frac{\log(d_A)}{n}$.

$\rho_{AB}$ is n-extendable if $\exists \tilde{\rho}^{AB_1 \dots B_n}$ s.t. $\rho^{AB} = \tilde{\rho}^{AB_i} \forall i$. Then we have $E_{sq}(\rho) \le \log(d_A)/n$.

We have that

$$E_D \le E_{sq} \le E_F$$

But if $E_{sq}(\rho) \le \epsilon^2$, is $\rho$ close to Sep? On PSET 9 we will show that if $\rho$ is $\frac{\log(d_A)}{\epsilon^2}$-extendable then we have that

$$\max_{M \in \text{1-LOCC}} \min_{\sigma \in \text{Sep}} |\text{tr}(M(\rho - \sigma))| \le \epsilon$$

This gives us non-trivial approximation bounds for runtime $d^{O(\log d)}$.