8.372 Quantum Information Science III (Fall 2026)

Lecturer: Aram W. Harrow
TA: Richard Allen
Scribe: Aarush Vailaya

Lectures 1–4 come from the 2024 lectures;
lectures 5 and onward come from 2026.

October 30, 2026

Contents

1. Bit commitment and purifications (2024) 1-1

2. Trace distance, fidelity, metrics (2024) 2-1

3. Classical information theory (2024) 3-1

4. Quantum compression (2024) 4-1

5. Quantum compression, relative entropy, hypothesis testing 5-1

6. Quantum relative entropy 6-1

7. Quantum sensing and Fisher information 7-1

8. Classical messages over quantum channels 8-1

9. HSW converse 9-1

10. Applications of the HSW converse 10-1

11. State tomography lower bound and resource inequalities 11-1

12. Randomness and concentration of measure 12-1

13. Random states, Rényi entropies, and entanglement 13-1

14. The symmetric subspace and Schur–Weyl duality 14-1

8.372 Quantum Information Science III **Fall 2024**

Lecture 1: September 5, 2024

Scribe: David D. Dai *Bit commitment and purifications*

1.1 Class Introduction

Topics

1. Quantum information theory and its mathematical foundations
2. Basic tools: norms, randomness, quantum entropies, and symmetry (group representations)
3. Applications: cryptography, many-body physics, optimization, and complexity / algorithms

Websites

1. Canvas: shell linking to everything else, email announcements
2. Piazza: discussion, questions (threaded conversations)
3. Gradescope: submit homework
4. Gitlab: lecture notes, homework problems
5. Overleaf: scribing

1.2 Information-Theoretically Secure Quantum Cryptography

Information-theoretically secure cryptography is secure against an adversary with infinite computational resources and time. This is stronger than security based on computational assumptions, such as RSA, which is based on the hardness of factoring. Some primitives that we might want to perform are:

1. Quantum key distribution: Alice and Bob want to share a secure random key and prevent eavesdropper Eve from learning the key. The goal is for Alice and Bob to finish the protocol with an identical key that Eve knows nothing about, or to abort.
2. Coin flipping: Alice and Bob are remote and need to simulate a fair coin flip. Letting the probability that the coin is 1 be p, there are two cases:
   - Strong: $p \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right]$ for some small ϵ no matter what. Alice and Bob cannot bias the coin in either direction.
   - Weak: Alice can bias $p \in \left[\frac{1}{2} - \epsilon, 1\right]$ and Bob can bias $p \in \left[0, \frac{1}{2} + \epsilon\right]$. This is useful if Alice prefers 0 and Bob prefers 1. For example, most people prefer to serve first in a sports match.
3. Oblivious transfer: Alice has a database $(x_0, x_1, x_2, \ldots)$, and Bob wants to access a specific value x_i. Bob doesn't want to reveal i to Alice, and Alice doesn't want to reveal the other values in the database to Bob.

4. Bit commitment: Alice writes a message, seals it in an envelope, and hands it to Bob (*commit* phase). Bob cannot read the message by himself (*hiding* property). Later, Alice can send instructions to Bob to reveal her earlier message (*reveal* phase). However, she cannot change the message after having committed it earlier (*binding* property). After the protocol concludes, Bob either learns the message (*valid* property) if nobody cheated, or he rejects.

Aside from quantum key distribution, all of these primitives have a similar trust model in which both parties are potentially honest or potentially adversarial. This situation is known as "two-party cryptography". Not all of these primitives are independent: oblivious transfer > bit commitment > strong coin flip > weak coin flip. Only weak coin flip and quantum key distribution are possible.

1.3 State Purification

The set of density matrices for a d-dimensional quantum system is:

$$D_d = \{\rho \in \mathcal{C}^{d\times d} : \rho \geq 0, \operatorname{Tr}\rho = 1\}, \tag{1.1}$$

where $\rho \geq 0$ means that ρ is positive semidefinite. Density matrices can be interpreted as a random ensemble of pure states, or as the marginal resulting from looking only at a small subsystem of a larger global pure state. In the marginal case, $\rho_A = \operatorname{Tr}_B(|\psi_{AB}\rangle\langle\psi_{AB}|)$, where $|\psi_{AB}\rangle$ is the global pure state, ρ_A is the density matrix for subsystem A, and Tr_B is the partial trace over the rest of the composite system.

We are interested in the inverse problem: given some fixed ρ_A, what is the set of all $|\psi_{AB}\rangle$ for which ρ_A is the reduced density matrix for subsystem A? Let d_A (d_B) be the dimension of subsystem A (B). Then the global pure states are:

$$|\psi_{AB}\rangle = \sum_{ij} C_{ij} |i\rangle \otimes |j\rangle, \quad \sum_{ij} |C_{ij}|^2 = 1. \tag{1.2}$$

The corresponding density matrix for subsystem A is:

$$\begin{aligned}
\rho_A &= \operatorname{Tr}_B(|\psi_{AB}\rangle\langle\psi_{AB}|) && (1.3)\\
&= \sum_k I \otimes \langle k| \sum_{ij} C_{ij} |i\rangle \otimes |j\rangle \sum_{i'j'} C_{i'j'} \langle i'| \otimes \langle j'| I \otimes |k\rangle && (1.4)\\
&= \sum_{ii'} \left[CC^\dagger\right]_{ii'} |i\rangle\langle i'|. && (1.5)\\
&= CC^\dagger && (1.6)
\end{aligned}$$

For all C, we can perform an SVD:

$$C = UDV^\dagger \to \rho_A = CC^\dagger = UD^2U^\dagger. \tag{1.7}$$

Except for the requirement that it be an isometry, V is unconstrained. D^2 is fixed because it corresponds to ρ_A's eigenvectors. U is also fixed up to rotations within eigenspaces, i.e. $U \to UR$ for unitary R such that $RD = DR$. Because R can be commuted through D as $(UR)DV^\dagger = UDRV^\dagger = UD(VR^\dagger)^\dagger$, the freedom in U can be folded into the freedom in V.

Therefore, the set of all purifications of ρ_A is:

$$\{\psi_{AB} = UDV^\dagger : V^\dagger V = I\}. \tag{1.8}$$

where D and U are fixed by the eigen-decomposition $\rho_A = UD^2U^\dagger$. The dimensions are:

$$\dim U = d_A \times r, \tag{1.9}$$
$$\dim D = r \times r, \tag{1.10}$$
$$\dim V = d_B \times r, \tag{1.11}$$

where r is the number of nonzero eigenvalues of ρ_A. For some V_1 with $\dim V_1 = d_{B_1} \times r$ and V_2 with $\dim V_2 = d_{B_2} \times r$, we can always find either an isometry W such that $V_2 = WV_1$ or $V_1 = WV_2$. To prove this, take $d_{B_2} \geq d_{B_1}$ WLOG. We can always complete the basis using Gram-Schmidt to create a $d_{B_1} \times d_{B_1}$ unitary $\tilde{V}_1$ which has V_1 as its first r columns. Additionally, use Gram-Schmidt to create $d_{B_2} \times d_{B_1}$ isometry $\tilde{V}_2$ that agrees with V_2 in its first r columns. Then the isometry between V_1 and V_2 is $W = \tilde{V}_2\tilde{V}_1^\dagger$. W clearly maps V_1 to V_2, and it is an isometry because $W^\dagger W = (\tilde{V}_2\tilde{V}_1^\dagger)^\dagger(\tilde{V}_2\tilde{V}_1^\dagger) = \tilde{V}_1\tilde{V}_2^\dagger\tilde{V}_2\tilde{V}_1^\dagger = I$.

We can also see that an isometry performed on subsystem B does not change ρ_A. In general, we have:

$$\begin{aligned} U \otimes V |\psi_{AB}\rangle &= \sum_{ij} C_{ij}\, U|i\rangle \otimes V|j\rangle\,, \\ &= \sum_{ii'jj'} U_{i'i} C_{ij} V_{j'j} |i'\rangle \otimes |j'\rangle\,, \\ &= \sum_{ij} \left[UCV^T\right]_{ij} |i\rangle \otimes |j\rangle\,. \end{aligned} \tag{1.12}$$

Since $\left(CV^T\right)\left(CV^T\right)^\dagger = C\left(V^\dagger V\right)^* C^\dagger = CC^\dagger$, $I \otimes V|\psi_{AB}\rangle$ and $|\psi_{AB}\rangle$ are purifications of the same ρ_A.

Putting the two directions together, we have the theorem: $|\psi_{AB}\rangle$ and $|\gamma_{AB'}\rangle$ purify the same density matrix ρ_A if and only if there exists some isometry W on the auxiliary spaces B and B' such that $I_A \otimes W|\psi_{AB}\rangle = |\gamma_{AB'}\rangle$ or $I_A \otimes W|\gamma_{AB'}\rangle = |\psi_{AB}\rangle$. The backward direction is very intuitive. Imagine that subsystem A is held by Alice on Earth, and subsystem B is held by Bob on Mars. By causality, an action that Bob takes alone is undetectable by Alice, i.e. cannot affect Alice's density matrix.

1.4 Proof that (Perfect) Bit Commitment is Impossible

There are three pictures of quantum operations: trace-preserving completely positive maps, Kraus operators, and isometries followed by partial traces. All are equivalent, and we use "isometry[1] followed by partial trace" here for convenience. We can actually ignore the partial trace: there is no difference between irreversibly throwing away the environment and merely not looking at it again. Ignoring the partial trace also allows for the possibility that a dishonest player may keep the environment and analyze it to gain an advantage instead of discarding it as instructed.

[1] An **isometry** V is a linear map from $\mathbb{C}^{d_A}$ to $\mathbb{C}^{d_B}$ for $d_B \geq d_A$ such that $V^\dagger V = I_{d_A}$. It is norm-preserving. Namely $\||\psi\rangle\| = \|V|\psi\rangle\|$, for any $|\psi\rangle \in \mathbb{C}^{d_A}$. One important fact is that $VV^\dagger = I_B$ iff $d_A = d_B$. A preliminary example of an isometry is to add a qubit state: $V : |\psi\rangle \mapsto |\psi\rangle \otimes |0\rangle$.
Moreoever, for a finite system (which is always what we consider), one can always extend the isometry V to an unitary $U : \mathbb{C}^{d_B} \to \mathbb{C}^{d_B}$ such that $U|v\rangle = W|v\rangle$ for any v in the domain of V. Then in such a case a quantum channel can be written as $\mathcal{E}(\rho) = \mathrm{Tr}_E(V\rho V^\dagger) = \mathrm{Tr}_E[U(\rho \otimes |\vec{0}\rangle\langle\vec{0}|_E)U^\dagger]$, where we have labeled the initial state of the environment before the action of the quantum channel as $|\vec{0}\rangle_E$.

Then after the commit phase, Alice and Bob share the global pure state $|\psi_{AB}^{(b)}\rangle$ out of the two choices $\{|\psi\rangle_{AB}^{(0)}, |\psi\rangle_{AB}^{(1)}\}$ for a committed (fixed) bit b. Because the protocol needs the hiding property, we have equation[2] $\psi_B^{(0)} = \psi_B^{(1)}$. Then by the above theorem, there exists some unitary U in Alice's Hilbert space such that $U \otimes I_B |\psi_{AB}^{(0)}\rangle = |\psi_{AB}^{(1)}\rangle$, which is not binding at all. Therefore, exact bit commitment is impossible.

[2]In this class we use the convention that a single Greek letter $\psi := |\psi\rangle \langle\psi|$ for a pure state $|\psi\rangle$, and ψ_B means we take the partial trace over the complement of subsystem B.

Lecture 2: September 10, 2024

Scribe: David D. Dai and Yeongwoo Hwang *Trace distance, fidelity, metrics*

2.1 Norms

2.1.1 General Properties

Norms measure "how big" an object is. They have three properties:

1. $\|cx\| = |c| \cdot \|x\| \quad \forall c \in \mathcal{C}$ (homogeneous)
2. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)
3. $\|x\| = 0 \leftrightarrow x = 0$, where 0 is the additive identity (separating)

If an operation satisfies the first two properties but not the third separating property, it is called a seminorm. A space equipped with a valid metric is a metric space.

2.1.2 Vector Norms

An important class of norms on vectors $\mathcal{C}^d$ are the L_p norms:

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}. \tag{2.1}$$

There are a few important special cases. The L^1 norm is the sum of the absolute values of the entries (Manhattan distance), the L^2 norm is the Euclidean norm, and the L^∞ norm is the maximum of the absolute values of the entries. Intuitively, p tells us how much the larger entries are weighed relative to the smaller entries. The L^∞ norm ignores all but the largest entry[1], while the L^1 norm treats all entries equally. Additionally, we require $p \geq 1$; $p < 1$ violates the triangle inequality, as can be seen easily for $x = (1, 0)$ and $y = (0, 1)$.

We call L_p and L_q *dual* if $1/p + 1/q = 1$. Then Hölder's inequality states:

$$\|x\|_p = \max_{\|y\|_q = 1} |\langle x, y \rangle| . \tag{2.2}$$

We will not prove Hölder's inequality but can inspect a few cases. $p = 2$ and $q = 2$ are dual, yielding:

$$\|x\|_2 = \max_{\|y\|_2 = 1} |\langle x, y \rangle| . \tag{2.3}$$

[1]Note that duplicated entries are not a problem. Even if there are m copies of the largest entry, the factor of m is suppressed by the $1/p$ power.

This is consistent with the Cauchy-Schwarz inequality: $\langle x, y\rangle \leq \sqrt{\langle x, x\rangle \langle y, y\rangle}$ with saturation if and only if $x \propto y$. $p = 1$ and $q = \infty$ are also dual, yielding:

$$\begin{aligned} \|x\|_1 &= \max_{\|y\|_\infty = 1} |\langle x, y\rangle|, \\ \sum_i |x_i| &= \max_{\max(|y_i|)=1} |\langle x, y\rangle|. \end{aligned} \tag{2.4}$$

This makes sense; if $x = (r_1 e^{i\theta_1}, r_2 e^{i\theta_2} \dots r_d e^{i\theta_d})$ for positive r and θ, then the maximum is achieved by $y = (e^{-i\theta_1}, e^{-i\theta_2} \dots e^{-i\theta_d})$. $p = \infty$ and $q = 1$ are also dual, yielding:

$$\begin{aligned} \|x\|_\infty &= \max_{\|y\|_1 = 1} |\langle x, y\rangle|, \\ \max(x_i) &= \max_{\sum_i |y_i| = 1} |\langle x, y\rangle|. \end{aligned} \tag{2.5}$$

This also makes sense; the maximum is achieved by $y_i = \delta_{i,\mathrm{argmax}(|x_i|)}$.

2.1.3 Matrix Norms

For some operator X, the Schatten p-norm S_p is:

$$\|X\|_{S_p} \equiv \|X\|_p = \|\Sigma(X)\|_p, \tag{2.6}$$

where $\Sigma(X)$ are the singular values of X. The Schatten p-norm of X is the L_p norm of X's singular values. All of the S_p norms have the nice property that they are invariant under left or right matrix multiplication by a unitary, because this does not change the singular values. If X is Hermitian, the singular values may be replaced with eigenvalues.

$S_\infty(X)$ corresponds to the maximum singular value of X, which is also the maximum factor by which X can stretch a vector by:

$$\|X\|_\infty = \max \Sigma(X) = \max_{\|v\|_2 = 1} \|Xv\|. \tag{2.7}$$

S_1 and S_2 can be expressed without using the SVD:

$$\|X\|_1 = \mathrm{Tr}\sqrt{X^\dagger X}, \|X\|_2 = \sqrt{\mathrm{Tr}\, X^\dagger X}, \tag{2.8}$$

where the square root is well-defined because $X^\dagger X$ is positive semi-definite. It is easy to show that Eq. 2.8 is consistent with Eq. 2.6 by plugging in $X = U\Sigma V^\dagger$.

We note that proving the triangle inequality for S_p is nontrivial.

2.1.4 Some Useful Sets

The unit sphere S and ball B with respect to some norm $\|\cdot\|$ are:

$$S = \{x : \|x\| = 1\}, \tag{2.9}$$

$$B = \{x : \|x\| \leq 1\}. \tag{2.10}$$

Below are a few sets commonly encountered in quantum information science:

- Pure quantum states: $S(L_2)$,
- Classical probability distributions: $S(L_1) \cap \{\text{nonnegative entries}\}$,
- Density matrices: $S(S_1) \cap \{\text{positive semidefinite}\}$,
- Measurement operators: $B(S_\infty) \cap \{\text{positive semidefinite}\}$.

2.2 Comparing Probability Distributions

2.2.1 Total Variation Distance

The total variation distance (TVD) between two probability distributions p and q is:

$$T(p,q) = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\sum_i |p_i - q_i|\,. \tag{2.11}$$

Note that

$$T(p,q) = \frac{1}{2}\sum_i |p_i - q_i| \le \frac{1}{2}\sum_i (p_i + q_i) = 1, \tag{2.12}$$

so $T(p,q) \in [0,1]$.

$T(p,q)$ has a nice operation definition. Consider a guessing game where we are given a random variable X, drawn either from distribution p or q with equal prior probability $1/2$. Given X's value, how often can we correctly guess which distribution it was drawn from? Bayes' rule gives the probability that X was drawn from p given $X = i$:

$$P(X \text{ from } p | X = i) = \frac{p_i}{p_i + q_i}. \tag{2.13}$$

The best strategy is to guess p if $P(X \text{ from } p|X=i) > 1/2$ and q otherwise, so the probability that we guess correctly given $X = i$ is

$$P(\text{correct}|X=i) = \max\left(\frac{p_i}{p_i+q_i}, \frac{q_i}{p_i+q_i}\right) = \frac{1}{2} + \frac{|p_i - q_i|}{2(p_i+q_i)}. \tag{2.14}$$

Then the probability that we guess correctly in general is:

$$\begin{aligned} P(\text{correct}) &= \sum_i P(\text{correct}|X=i)P(X=i) \\ &= \sum_i \left[\frac{1}{2} + \frac{|p_i - q_i|}{2(p_i+q_i)}\right]\left[\frac{p_i+q_i}{2}\right] \\ &\frac{1}{2} + \frac{T(p,q)}{2}. \end{aligned} \tag{2.15}$$

For example, if $T(p,q) = 1/2$, then we can correctly guess whether a random variable was drawn from p or q three-quarters of the time.

2.2.2 Fidelity (Bhattacharyya Coefficient)

An alternative way of comparing probability distributions is the fidelity:

$$F(p,q) = \langle \sqrt{p}, \sqrt{q} \rangle \,, \tag{2.16}$$

where $\sqrt{p}$ is the element-wise square root of the vector of probabilities. The square root is necessary to guarantee that $F(p,p) = 1$ for all p, something which would not be true for $\langle p, p \rangle$.

If our random variable comes from concatenating two independent random variables, i.e. $i = (a,b), p_i = p_a p_b$, then the fidelity factorizes:

$$\begin{aligned} F(p_i, q_i) &= \sum_i \sqrt{p_i q_i} \\ &= \sum_{a,b} \sqrt{p_a p_b q_a q_b} \\ &= F(p_a, q_a) F(p_b, q_b). \end{aligned} \tag{2.17}$$

The TVD notably lacks this property. The total variation distance and fidelity also satisfy the inequalities

$$1 - F \leq T \leq \sqrt{2(1-F)}. \tag{2.18}$$

Even through T doesn't factorize, Eq. 2.18 allows us to bound $T(p^{\otimes n}, q^{\otimes n})$, where $p^{\otimes n}$ means the probability distribution corresponding to drawing n times from p. In particular, T approaches 1 exponentially in n.

2.3 Quantum Distinguishability

What is the appropriate metric via which we should compare quantum states? A first guess could be the ℓ_2 vector norm, i.e. $\| \, |p\rangle - |q\rangle \, \|_2$. This is equal to $\sqrt{2(1 - \mathrm{Re}(\langle p|q\rangle)}$ and has an undesirable sensitivity to relative phase. By maximizing over the global phase, we obtain the closeness measure,

$$| \, \langle p|q\rangle \, |$$

which we'll define as the *fidelity* between $|p\rangle$ and $|q\rangle$. However, this definition also has a drawback, which is that there is no nice "operational" interpretation of fidelity. For that, we introduce the *trace distance*

Definition 2.3.1 (Trace Distance)**.** *Let ρ, σ be two mixed states. The trace distance between ρ, σ is denoted $T(\rho, \sigma)$ and is defined equivalently as,*

$$T(\rho,\sigma) = \frac{1}{2} \|\rho - \sigma\|_1 \tag{2.19}$$

$$= \max_{0 \preceq M \preceq \mathbb{I}} \mathrm{tr}[M(\rho - \sigma)] \tag{2.20}$$

This metric has some nice properties,

- (Unitary Invariance) $T(V\rho V^\dagger, V\sigma V^\dagger) = T(\rho,\sigma)$
- (Data Processing Inequality or Monotonicity) $T(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leq T(\rho,\sigma)$

In fact, by defining our channel via the measurement obtaining the maximum in (2.20) we can saturate the monotonicity bound:

$$\mathcal{E}(\rho) := \text{tr}[M\rho]\,|0\rangle\langle 0| + \text{tr}[(\mathbb{I} - M)\rho]\,|1\rangle\langle 1|$$

Note that we've define trace distance over mixed states, whereas our definition of fidelity was limited to pure states. We can generalize to mixed states as follows,

Definition 2.3.2 (Fidelity)**.** *Let ρ, σ be two mixed states. The fidelity between ρ, σ is denoted $F(\rho, \sigma)$ and is defined as*

$$F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1 = \text{tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]$$

Some nice properties of fidelity are,

- (Fuchs-van de Graaf Inequalities) $1 - F \leq T \leq \sqrt{1 - F^2}$
- (DPI or Monotonicity) $F(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leq F(\rho, \sigma)$

We'll prove the Fuchs-van de Graaf inequalities in the problem set. Interestingly, fidelity *does not* satisfy the triangle inequality (and thus is not a metric); however, $\arccos(F(\cdot, \cdot))$ does.

2.3.1 Uhlmann's Theorem

The definition we gave for fidelity is quite cumbersome; in practice it can be very annoying to compute the square roots of matrices. Uhlmann's theorem gives a nice alternative characterization of the fidelity between mixed states.

Theorem 2.3.1 (Uhlmann's Theorem)**.** *Let ρ, σ be two mixed states defined over a quantum register A. Then,*

$$F(\rho, \sigma) = \max_{\substack{|\rho\rangle_{AB} \text{ s.t. } \text{tr}_B[|\rho\rangle\langle\rho|]=\rho \\ |\sigma\rangle_{AB} \text{ s.t. } \text{tr}_B[|\sigma\rangle\langle\sigma|]=\sigma}} |\,\langle\rho|\sigma\rangle\,|$$

i.e. the mixed state fidelity between ρ and σ is the maximum pure state fidelity between purifications of ρ and σ.

Before giving the proof of this theorem, we define the "canonical" purification of a mixed state. To do so, we define the (unnormalized) maximally entangled state between two registers A and B of equal dimension d as,

$$|\Gamma\rangle = \sum_{i=1}^{d} |i\rangle_A\,|i\rangle_B$$

Definition 2.3.3 (Canonical Purification)**.** *For a mixed state ρ, its canonical purification is denoted by $|\phi^\rho\rangle$ and defined as,*

$$|\phi^\rho\rangle := (\sqrt{\rho}_A \otimes \mathbb{I}_B)\,|\Gamma\rangle_{AB}$$

The fact that $\text{tr}_B[\phi^\rho] = \rho$ can be verified by a simple computation. We'll also need the following lemma, which, intuitively, we should think of as the matrix analogue of "tuning" the phases of a probability distribution to obtain the ℓ_1 norm.

Lemma 2.3.1.

$$\max_U |\,\text{tr}[AU]| = \|A\|_1$$

Proof. Take the singular-value decomposition of A to obtain $A = UDW^\dagger$. Then, $\mathrm{tr}[AU] = \mathrm{tr}\left[DW^\dagger UV\right]$, where we've used the cyclic property of the trace. Rather than maximizing over U, consider maximizing over $U = W\tilde{U}V^\dagger$. This is equivalent as for an original $U^\star$, we can set $\tilde{U} = W^\dagger U^\star V$ and then $U = U^\star$. Thus,

$$\max_U \mathrm{tr}\left[DW^\dagger UV\right] = \max_{U = W\tilde{U}V^\dagger} \mathrm{tr}\left[DW^\dagger UV\right] = \max_{\tilde{U}} \mathrm{tr}\left[D\tilde{U}\right]$$

But since D is a diagonal matrix, the RHS is just $\max_U \sum_i D_{i,i}U_{i,i} \leq \mathrm{tr}[|D|] = \|A\|_1$. □

We now give the proof of Uhlmann's theorem.

Proof of Theorem 2.3.1. Recall that all purifications of a mixed state ρ_A as $|\rho\rangle_{AB}$ are equivalent under a unitary on just the B register. As a result, we can replace $\max_{|\rho\rangle,|\sigma\rangle} |\langle\rho|\sigma\rangle|$ with

$$\max_{U,V} |\langle\phi^\rho| (\mathbb{I}\otimes U_B)(\mathbb{I}\otimes V_B) |\phi^\sigma\rangle| \tag{2.21}$$

But U_BV_B is just another unitary and can think of this as fixing $|\phi^\rho\rangle$ and only maximizing over a single unitary (which is equivalent to maximizing over purifications of σ). Then,

$$\begin{aligned}
(2.21) &= \max_U |\langle\phi^\rho| (\mathbb{I}\otimes U) |\phi^\sigma\rangle| && (2.22)\\
&= \max_U \langle\Gamma| (\sqrt{\rho}\otimes\mathbb{I})(\mathbb{I}\otimes U)(\sqrt{\sigma}\otimes\mathbb{I}) |\Gamma\rangle && (2.23)\\
&= \max_U \langle\Gamma| (\sqrt{\rho}\sqrt{\sigma})\otimes U |\Gamma\rangle && (2.24)\\
&= \max_U \mathrm{tr}\left[\sqrt{\rho}\sqrt{\sigma}U^\top\right] && (2.25)\\
&= \|\sqrt{\rho}\sqrt{\sigma}\|_1 && (2.26)
\end{aligned}$$

where in (2.25) we've used that the maximally mixed state over registers A, B satisfies $(\mathbb{I}\otimes U)|\Gamma\rangle = (U^\top\otimes\mathbb{I})|\Gamma\rangle$. The last equality uses Lemma 2.3.1. □

2.4 No-go Theorem for Bit Commitment

To conclude the lecture, we revisit the no-go theorem from the first lecture and relax the hiding condition so that Bob is allowed some small probability of recover Alice's commitment. Formally, let's say that an honest Alice commits to the states $|\psi_0\rangle_{AB}$ or $|\psi_1\rangle_{AB}$. Then, a limit on Bob's ability to distinguish these two states corresponds to requiring,

$$T(\mathrm{tr}_A[\psi_0], \mathrm{tr}_A[\psi_1]) \leq \varepsilon \implies F(\mathrm{tr}_A[\psi_0], \mathrm{tr}_A[\psi_1]) \geq 1-\varepsilon$$

Since $|\psi_0\rangle_{AB}$ and $|\psi_1\rangle_{AB}$ are purifications of $\mathrm{tr}_A[\psi_0]$ and $\mathrm{tr}_A[\psi_1]$, we know that there exists a unitary U such that,

$$F((U\otimes\mathbb{I})|\psi_0\rangle, |\psi_1\rangle) \geq 1-\varepsilon$$

Define $|\psi_{\text{fake}}\rangle := (U\otimes\mathbb{I})|\psi_0\rangle$. Converting back to trace distance, we have that

$$T(\psi_{\text{fake}}, \psi_1) \leq \sqrt{2\varepsilon} \overset{\text{monotonicitiy}}{\implies} T(\mathcal{E}_{\text{reveal}}(\psi_{\text{fake}}), \mathcal{E}_{\text{reveal}}(\psi_1)) \leq \sqrt{2\varepsilon}$$

We conclude that Bob cannot distinguish between $|\psi_{\text{fake}}\rangle$, which corresponds to $|\psi_0\rangle$ with a unitary applied to only Alice's side, and $|\psi_1\rangle$. Thus, this protocol is not binding.

8.372 Quantum Information Science III **Fall 2024**

Lecture 3: September 12, 2024

Scribe: Aditi Venkatesh and Jin Ming Koh *Classical information theory*

3.1 Introduction

There are parallels between classical and quantum information theory.

Application	**Classical information theory**	**Quantum information theory**
Data compression	Shannon entropy	von Neumann entropy
Channel coding	Mutual information	Quantum mutual information (for *classical* information over noisy channel); coherent information (for *quantum* information over noisy channel).
Hypothesis testing	Relative entropy	Quantum relative entropy

3.2 Entropy

Entropy is a measure of uncertainty.

3.2.1 Shannon entropy

Definition 3.2.1 (Shannon entropy of single variable)**.** *For random variable $X \sim p$ such that $\mathbb{P}(x) = p(x)$, the Shannon entropy of X is*

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{3.1}$$

Some properties:

1. *Bounds.* The Shannon entropy satisfies $0 \leq H(X) \leq \log_2 d$ for d the size of the alphabet of X. The lower bound is attained for deterministic X. The upper bound is attained for uniformly random X.

2. *Concavity.* For all $0 \leq \lambda \leq 1$,

$$\lambda H(p_1) + (1-\lambda) H(p_2) \leq H[\lambda p_1 + (1-\lambda) p_2]. \tag{3.2}$$

3. *Norm power series expansion.*

$$\|p\|_{1+\epsilon} = \left(\sum_{x \in X} p(x)^{1+\epsilon} \right)^{\frac{1}{1+\epsilon}} = 1 + \epsilon H(p) + \mathcal{O}\big(\epsilon^2\big). \tag{3.3}$$

Definition 3.2.2 (Shannon entropy of two variables)**.** *For random variables $(X, Y) \sim p$ such that $\mathbb{P}(x, y) = p(x, y)$, the Shannon entropy of the joint distribution*

$$H(X, Y) = -\sum_{(x,y)\in(X,Y)} p(x, y) \log_2 p(x, y). \tag{3.4}$$

For a product distribution, $H(X, Y) = H(X) + H(Y)$.

3.2.2 Conditional entropy

Definition 3.2.3 (Conditional entropy)**.** *For random variables $(X, Y) \sim p$ such that $\mathbb{P}(x, y) = p(x, y)$, the conditional entropy of Y given X is*

$$\begin{aligned} H(Y|X) &= -\sum_{x\in X} \mathbb{P}(X = x) H(Y|X = x) && (3.5)\\ &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log p(y|x) && (3.6)\\ &= H(X, Y) - H(X), && (3.7) \end{aligned}$$

where we have noted $p(y|x) = p(x, y)/p(x)$.

The physical intuition is that $H(X)$ is the uncertainty of X, whereas $H(Y|X)$ is the uncertainty of Y when we know X. Therefore $H(X, Y) = H(X) + H(Y|X)$.

Remark 3.2.1 (Chain rule)**.** *For random variables (X_1, X_2, X_3),*

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2). \tag{3.8}$$

Remark 3.2.2 (Non-negativity of conditional entropy)**.** *Classically, $H(Y|X) \geq 0$. But not so quantumly. An example is an EPR pair. Then $H(X, Y) = 0$ as the two subsystems jointly are in a pure state, but $H(X) = H(Y) = 1$ as the reduced density matrix of each subsystem is maximally mixed. That is, quantumly, the joint probability distribution can possess less entropy than its marginal distributions.*

3.2.3 Mutual information

Definition 3.2.4 (Mutual information)**.** *The mutual information between random variables (X, Y) is*

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{3.9}$$

The physical interpretation is that $I(X; Y)$ is how much information one learns about X when one looks at Y, or symmetrically, how much information one learns about Y when one looks at X.

Remark 3.2.3 (Non-negativity of mutual information)**.** *Both classically and quantumly,*

$$I(X; Y) \geq 0 \iff H(X) + H(Y) \geq H(X, Y) \iff H(Y|X) \leq H(Y). \tag{3.10}$$

3.2.4 Entropy of density matrices

Definition 3.2.5 (Shannon entropy of density matrix)**.** *The Shannon entropy of a density matrix ρ is*

$$H(\rho) = -\sum_k \lambda_k \log_2 \lambda_k \tag{3.11}$$

where $\{\lambda_k\}_k$ are the eigenvalues of the matrix.

The entropy $H(\rho)$ is the Shannon entropy of the probability distribution of measurement outcomes obtained when ρ is measured in its eigenbasis.

Remark 3.2.4 (Strong sub-additivity)**.** *For quantum systems A, B and C,*

$$H(A) + H(ABC) \leq H(AB) + H(AC). \tag{3.12}$$

3.3 Noiseless coding theorem

Theorem 3.3.1 (Noiseless coding theorem)**.** *It is possible to compress n length iid message $x_1, x_2, \ldots, x_n$, from x X, to $nH(X) + o(1)$ bits, with perfect recovery.*

Proof. Lets consider the probability distribution $X := \{x, p(x)\}$ where each letter x_i has probability $p(x_i)$. For an n-letter message,

$$p(x_1 x_2 ... x_n) = \prod_{i=1}^{n} p(x_i)$$

due to iid. Unless X is uniformly random it is possible to compress this distribution to an smaller string. Using the law of large numbers we know that for a string of n letters, x_i typically occurs $np(x_i)$ times. Therefore using Stirling's approximation we can say that the number of typical strings is

$$\frac{n!}{\prod_x (np(x))!} \approx 2^{nH(X)}$$

where,

$$H(X) = -\sum p(x) \log_2 p(x)$$

If we use a block code that relates integers to typical sequences of the n-letter message, then the information in the n-letter string can be conveyed in on average $nH(X)$ bits. We need the $+o(1)$ in order to prove achievability. □

3.4 Noisy coding theorem

Consider now that the channel over which we transmit information is noisy. We encode our input message, pass the encoded message over the channel, and decode at the destination.

Definition 3.4.1 (Rate)**.** *Using a message of length n sent over the channel to encode a message of length k, the rate of the transmission is $R = k/n$.*

Definition 3.4.2 (Channel capacity)**.** *Consider a noisy channel which receives random variable X as input and outputs random variable Y. Then the channel capacity is*

$$C = \max_X I(X; Y), \tag{3.13}$$

where the maximization is performed over the input random variable X.

Theorem 3.4.1 (Noisy coding theorem)**.** *Consider sending a message of length n over a noisy channel at rate R. It is possible to do so with vanishing probability of error as $n \to \infty$ as long as $R < C$ where C is the capacity of the channel. Otherwise, the probability of error approaches unity as $n \to \infty$.*

Proof. Consider a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The channel is characterized by a conditional probability distribution $P(y|x)$, which gives the probability of receiving symbol $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ is sent from Alice to Bob. R and C of the code are defined as

$$R = \frac{\log_2 M}{n},$$
$$C = \max_{P(x)} I(X;Y),$$

where M is the number of codewords, and n is the length of each codeword. Construct a random codebook by selecting $M = 2^{nr}$ codewords $x_1, x_2, \ldots, x_M$ independently and uniformly from $\mathcal{X}^n$ according to the distribution $P(x)$. Bob observes the output $y^n \in \mathcal{Y}^n$ and decodes the received message to one of the M possible codewords. The goal is to show that for $r < C$, the probability of error can be made arbitrarily small as $n \to \infty$. Define the jointly typical set $T_\epsilon^{(n)}(X,Y)$ as the set of pairs (x^n, y^n) such that:

$$\left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon,$$
$$\left| -\frac{1}{n} \log P(y^n) - H(Y) \right| < \epsilon,$$
$$\left| -\frac{1}{n} \log P(x^n, y^n) - I(X;Y) \right| < \epsilon.$$

Decoding is performed by finding the unique codeword x_i^n such that the pair (x_i^n, y^n) is jointly typical.

The probability of error can be decomposed into two types: 1. No codeword x_i^n is jointly typical with y^n. 2. There exists a codeword x_j^n (with $j \neq i$) that is jointly typical with y^n.

The probability of the first type of error vanishes as $n \to \infty$, by the law of large numbers and the properties of typical sets. For the second type of error, using the union bound and the independence of codewords, we get:

$$P(\text{error}) \leq P(\text{incorrect decoding}) \leq (M-1)P(\text{codeword typical with } y^n).$$

Since the number of codewords $M = 2^{nr}$ and the probability that a randomly chosen codeword is jointly typical with y^n is approximately $2^{-nI(X;Y)}$, the probability of error is bounded by:

$$P(\text{error}) \leq (M-1)2^{-nI(X;Y)} \approx 2^{n(R-I(X;Y))}.$$

Thus, for $R < C$, the probability of error tends to zero as $n \to \infty$. Conversely, if $r > C$, the probability of error approaches 1 as $n \to \infty$.

Therefore, reliable communication over a noisy channel is possible at any rate $R < C$, and the probability of error can be made arbitrarily small. Conversely, for rates $R > C$, the probability of error approaches 1.

□

Lecture 4: September 17, 2024

Scribe: Aditi Venkatesh *Quantum Compression*

4.1 Classical Compression

Last class we talked about Shannon's noiseless coding theorem and data compression. We said if $R > H(p)$ compression is possible (direct) and if $R < H(p)$ compression is impossible (converse). Here we will prove the converse (the direct proof is in the previous lecture note). We can define a typical set as

$$T_{p,\delta}^n = \{x^n = (x_1, x_2, ..., x_n) : \left| -\frac{1}{n} \log p^{\otimes n}(x^n) - H(p) \right| < \delta\}$$

$$\epsilon = 1 - p^{\otimes n}(T_{p,\delta}^n) \leq 2^{-n\delta} \to 0 \quad \text{as } n \to \infty$$

This implies that the set of non-typical sequences becomes vanishingly small as n increases.

Now, suppose we want to compress the source X^n to k bits, where $k < nH(p)$. The goal is to show that this compression will lead to a vanishing probability of correctly decoding the original sequence.

Consider the compression process:

$$\boldsymbol{x}^n \xrightarrow{\text{Encoder (E)}} m \xrightarrow{\text{Decoder (D)}} \boldsymbol{x}^n$$

Here, $\boldsymbol{x}^n$ is the original sequence, m is the compressed version (with k bits), and $\boldsymbol{x}^n$ is the decoded sequence. The encoder uses r as a random seed to perform the compression. Pick r to maximize

$$\mathbb{P}\left(D(E(\boldsymbol{x}^n)) = \boldsymbol{x}^n \mid r\right)$$

Let $S \subseteq \Sigma^n$ be the set of sequences that can be decoded correctly. Since we're compressing to k bits, the size of this set is constrained by $|S| \leq 2^k$. We now want to bound the probability that a sequence $\boldsymbol{x}^n$ is decoded correctly.

$$p^n(S) \leq p^n(T_{p,\delta}^n \cap S) + p^n(T_{p,\delta}^{n^c})$$

The second term, $p^n(T_{p,\delta}^{n^c})$, is the probability of being in the non-typical set, which is upper-bounded by ϵ, a small quantity that goes to zero as $n \to \infty$. The first term can be bounded using the size of S and the fact that typical sequences occur with probability around $2^{-nH(p)}$:

$$p^n(T_{p,\delta}^n \cap S) \leq 2^k 2^{-nH(p)+n\delta}$$

Thus, the total probability is:

$$p^n(S) \leq 2^k 2^{-nH(p)+n\delta} + \epsilon$$

$$\to 0 \quad \text{if} \quad \frac{k}{n} < H(p)$$

This shows that if the compression rate k/n is less than $H(p)$, the probability of correctly decoding the sequence goes to zero as n increases. Therefore, it is not possible to compress below the entropy rate without losing information.

4.2 Quantum Compression (Entropy)

The entropy of a density matrix ρ, also known as the von Neumann entropy, is defined as:

$$S(\rho) = H(\text{eig}(\rho)) = -\text{Tr}(\rho \log \rho),$$

where $H(\text{eig}(\rho))$ is the Shannon entropy of the eigenvalues of ρ.

4.2.1 Bounds on Entropy

The von Neumann entropy satisfies the following bounds:

$$0 \leq S(\rho) \leq \log d,$$

where d is the dimension of the Hilbert space.

If $S(\rho) = 0$, then the eigenvalues of ρ are $(1, 0, 0, \ldots, 0)$, implying that $\rho = |\psi\rangle\langle\psi|$ for some pure state $|\psi\rangle$. If $S(\rho) = \log d$, then ρ is the maximally mixed state:

$$\rho = \frac{I}{d}.$$

4.2.2 Conditional Entropy

The entropy of a system X is given by:

$$S(X) = S(\rho_X).$$

The conditional entropy is defined as:

$$S(X|Y) = S(XY) - S(Y),$$

which can be negative. This definition carries over from classical information theory.

4.2.3 Mutual Information

The mutual information between two systems X and Y is defined as:

$$I(X : Y) = S(X) + S(Y) - S(XY),$$

which can also be expressed as:

$$I(X : Y) = S(X) - S(X|Y).$$

4.3 Typical Subspaces and Projectors

4.3.1 Definition of Typical Subspace

Let ρ be a density matrix acting on a Hilbert space $\mathcal{H}$, and let $\epsilon > 0$ be a small positive number. The typical subspace, denoted $\mathcal{T}_\epsilon^{(n)}$, is defined as the span of the eigenvectors of $\rho^{\otimes n}$ corresponding to eigenvalues close to $2^{-nS(\rho)}$, where $S(\rho)$ is the von Neumann entropy of ρ:

$$S(\rho) = -\text{Tr}(\rho \log \rho).$$

The typical subspace satisfies the following properties:

- $\mathbb{P}\left(\psi \in \mathcal{T}_\epsilon^{(n)}\right) \geq 1 - \epsilon$ for a random state ψ drawn from $\rho^{\otimes n}$.
- The dimension of the typical subspace is approximately $2^{nS(\rho)}$ for large n.

4.3.2 Properties of Typical Subspaces

1. *High Probability Support*: A quantum state ψ drawn according to $\rho^{\otimes n}$ has a high probability of being in the typical subspace. This is crucial for understanding the structure of quantum information over many copies.

2. *Dimensionality*: The typical subspace has dimension $d_{\text{typ}} \approx 2^{nS(\rho)}$, where $S(\rho)$ is the von Neumann entropy of ρ. This shows that the size of the subspace grows exponentially with the number of copies n.

4.4 Projectors onto Typical Subspaces

Given the typical subspace $\mathcal{T}_\epsilon^{(n)}$, we define a projector P_{typ} that projects any state onto this subspace.

4.4.1 Construction of the Projector

Let $\rho^{\otimes n}$ be the n-fold tensor product of the density matrix ρ. We diagonalize $\rho^{\otimes n}$ in its eigenbasis:

$$\rho^{\otimes n} = \sum_i \lambda_i |i\rangle\langle i|.$$

The typical subspace corresponds to the eigenvalues λ_i that satisfy

$$2^{-n(S(\rho)+\delta)} \leq \lambda_i \leq 2^{-n(S(\rho)-\delta)},$$

where $\delta > 0$ is a small positive number. The projector onto the typical subspace is given by

$$P_{\text{typ}} = \sum_{i:\lambda_i \text{ typical}} |i\rangle\langle i|.$$

4.4.2 Properties of the Projector

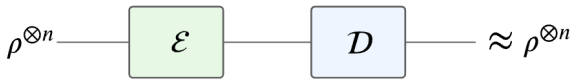
The projector P_{typ} has the following important properties:

- *Approximate Preservation of Trace*: For large n, we have $\text{Tr}(P_{\text{typ}}\rho^{\otimes n}) \geq 1 - \epsilon$. This means that most of the probability mass of $\rho^{\otimes n}$ lies within the typical subspace.
- *Dimensionality*: The rank of P_{typ} (the dimension of the typical subspace) is approximately $2^{nS(\rho)}$.

4.5 Quantum Compression

Below we give four possible schemes of quantum compression and decide whether or not they are valid.

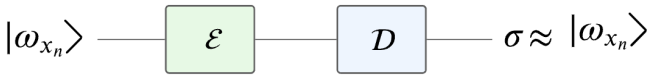
1. In the first scheme we have n copies of a density matrix ρ and we apply an encoder $\mathcal{E}$ and then a decoder $\mathcal{D}$ and check if the final density matrix matches the initial one.



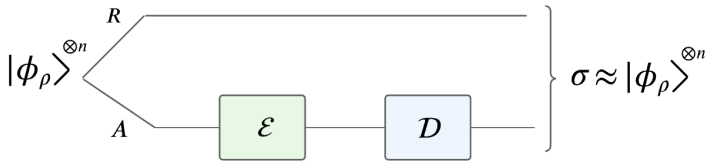
2. In the second scheme we can define x_n λ_n and check $\mathop{\mathbb{E}}_{x^n \sim \lambda^n} F(\sigma, |v_{x_n}\rangle) \approx 1$.



3. In the third scheme we define $\rho = \sum_i p_i |\omega_i\rangle \langle\omega_i|$ and check if the final state is close to the initial state.



4. In the fourth scheme we have n copies of a state $|\phi_\rho\rangle$ which we entangle with a reference. Then we apply the encoder and the decoder and check if the joint state is close to the initial state.



The first scheme does not work because it does not preserve correlations. We can see that the fourth, third, and second are equivalent ($4 \to 3 \to 2 \to 4$). It is easy to show that ($3 \to 2$) by choosing the eigenbasis. To show ($4 \to 3$), we use the fact that $\forall$ ensembles $\{p_i, |\omega_i\rangle\}$ such that $\rho = \sum_i p_i |\omega_i\rangle \langle\omega_i|$ there exists measurement operators $M_1, ..., M_l$ on R that induces this ensemble.

Lecture 5: September 24, 2026

Scribe: Aarush Vailaya *Quantum compression, relative entropy, hypothesis testing*

5.1 Quantum compression

Definition 5.1.1 (Naive compression)**.** *A compression scheme takes*

$$\rho^{\otimes n} \xrightarrow{\mathcal{E}} k \text{ qubits} \xrightarrow{\mathcal{D}} \sigma \approx \rho^{\otimes n}. \tag{5.1}$$

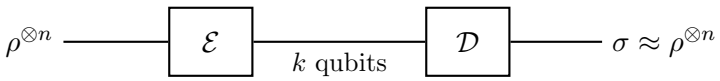


Figure 5.1: The naive compression scheme of Definition 5.1.1: n copies of ρ are encoded into k qubits and then decoded. The rate is $R = k/n$.

This definition is bad; let us see why. Here is a review of the classical case.

Let P be a distribution on $[d] = \{1, \dots, d\}$, and let $x^n \sim P^n$. The encoding is a map $E : [d]^n \to \{0,1\}^k$, where $R = k/n$ is the rate of the encoding. The decoding map is $D : \{0,1\}^k \to [d]^n$. We define the error rate

$$\epsilon = \Pr\big[D(E(x^n)) \neq x^n\big]. \tag{5.2}$$

Shannon's noiseless coding theorem says that the smallest achievable rate (with $\epsilon \to 0$ as $n \to \infty$) is $H(P)$.

Let P be uniformly distributed over one bit. It has entropy 1, so n bits can only be compressed to n bits.

Now look at the quantum case; there is a little issue with the diagram. Suppose the pure states we want to transmit are truly randomly distributed, so that the average state ρ is the maximally mixed state I/d (locally, this looks like half of a maximally entangled state). Then, say the compression scheme throws away our qubits and just outputs the maximally mixed state $(I/d)^{\otimes n}$. The output distribution is correct, since it equals $\rho^{\otimes n}$ exactly, but it is not correlated at all with the input. So Definition 5.1.1 is satisfied with $k = 0$ qubits, which is absurd.

Essentially, in the classical case we take a probability over all strings (the input is a string, and we must recover that string). In the quantum case above, we are implicitly averaging all our states into a single density matrix ρ, so we lose information in the description above.

Definition 5.1.2 (Eigenbasis fidelity)**.** *Write $\rho = \sum_x \lambda_x |v_x\rangle\langle v_x|$ in its eigenbasis, and input eigenvectors into the scheme:*

$$|v_{x^n}\rangle \xrightarrow{\mathcal{E}} \xrightarrow{\mathcal{D}} \sigma(x^n). \tag{5.3}$$

We demand

$$\mathbb{E}_{x^n \sim \lambda^n} F\big(|v_{x^n}\rangle\langle v_{x^n}|, \sigma(x^n)\big) \approx 1. \tag{5.4}$$

Essentially, each eigenvector of $\rho^{\otimes n}$ should be mapped with relatively high fidelity.

Definition 5.1.3 (Ensemble fidelity)**.** *A stronger way to state this: if $\rho = \sum_i p_i |w_i\rangle\langle w_i|$ with a not-necessarily orthogonal set of states $|w_i\rangle$, we want*

$$\mathbb{E}_{i^n \sim p^n} F\left(|w_{i^n}\rangle\langle w_{i^n}|, \mathcal{D}(\mathcal{E}(|w_{i^n}\rangle\langle w_{i^n}|))\right) \approx 1 \tag{5.5}$$

for every such ensemble decomposition of ρ.

This is a stronger condition than Definition 5.1.2 (which is the special case where the ensemble is the eigen-decomposition), as it says that no matter what distribution of pure states we think about, we would like to be right on average.

Definition 5.1.4 (Entanglement fidelity)**.** *Let $|\phi_\rho\rangle$ be a purification of ρ on $R \otimes A$ (R is the reference system, A is our system), and define*

$$\sigma = (I_R \otimes \mathcal{D} \circ \mathcal{E})\left(|\phi_\rho\rangle\langle\phi_\rho|^{\otimes n}\right). \tag{5.6}$$

We want $F\left(|\phi_\rho\rangle\langle\phi_\rho|^{\otimes n}, \sigma\right) \approx 1$.

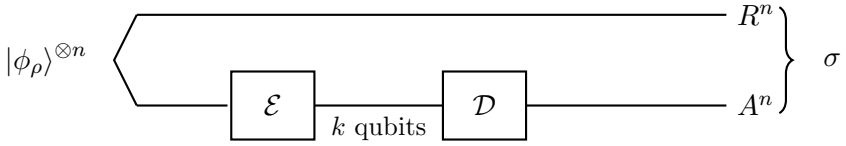


Figure 5.2: Entanglement fidelity (Definition 5.1.4): the reference system R^n is left untouched while A^n goes through the compression scheme. We ask that the joint output σ be close to the original purification.

It turns out Definition 5.1.4 implies Definition 5.1.3. In fact, Definitions 5.1.3 and 5.1.4 are practically equivalent (where the error for the ≈ 1 may be a bit lossy in conversion), while Definition 5.1.2 is not good.

Consider

$$\rho = \begin{pmatrix} 2/3 & 0 \\ 0 & 1/3 \end{pmatrix}. \tag{5.7}$$

One decomposition is into the states $|w_\pm\rangle = \sqrt{2/3}\,|0\rangle \pm \sqrt{1/3}\,|1\rangle$, each with probability $1/2$.

Consider the following measure-and-prepare scheme. The encoder measures each qubit in the eigenbasis $\{|0\rangle, |1\rangle\}$ of ρ, obtaining a string x^n, and compresses x^n with a classical code at rate $H(\lambda) = S(\rho)$. The decoder recovers x^n and prepares the product state $|x^n\rangle = |x_1\rangle \otimes \cdots \otimes |x_n\rangle$. Note this uses the same rate as Schumacher compression.

Eigenvectors are preserved faithfully so this satisfies Definition 5.1.2, but on a single copy of $|w_+\rangle$, the measurement gives 0 with probability $2/3$ and 1 with probability $1/3$, so the output is $\sigma = \frac{2}{3}|0\rangle\langle 0| + \frac{1}{3}|1\rangle\langle 1|$ and

$$F(|w_+\rangle\langle w_+|, \sigma) = \langle w_+|\,\sigma\,|w_+\rangle = \tfrac{2}{3}\cdot|\langle w_+|0\rangle|^2 + \tfrac{1}{3}\cdot|\langle w_+|1\rangle|^2 = \left(\tfrac{2}{3}\right)^2 + \left(\tfrac{1}{3}\right)^2 = \tfrac{5}{9}, \tag{5.8}$$

and likewise for $|w_-\rangle$. Over n copies the fidelities multiply, so the fidelity on any $|w_{i^n}\rangle$ is $(5/9)^n \to 0$.

Theorem 5.1.1 (Schumacher–Jozsa compression)**.** *For any $\delta > 0$, there is a compression scheme using $k \approx n(S(\rho) + \delta)$ qubits achieving entanglement fidelity $\to 1$ as $n \to \infty$. Conversely, any rate $R < S(\rho)$ has fidelity bounded away from 1.*

Proof of achievability. With typicality from before, the typical projector $\Pi^n_{\rho,\delta}$ satisfies

$$\operatorname{tr}\left[\Pi^n_{\rho,\delta}\rho^{\otimes n}\right] \geq 1-\epsilon, \qquad \operatorname{tr}\Pi^n_{\rho,\delta} \leq 2^{n(S(\rho)+\delta)}. \tag{5.9}$$

The idea is that we project onto a subspace of our space, and we want to rotate this subspace to be just the first roughly $n(S(\rho)+\delta)$ qubits; then we can discard the rest (if the measurement $\{\Pi^n_{\rho,\delta}, I-\Pi^n_{\rho,\delta}\}$ fails, the encoder outputs some fixed state.) This is pretty mechanical, what is interesting is fidelity. We denote

$$|\psi\rangle = \frac{(\Pi^n_{\rho,\delta}\otimes I)\,|\phi_\rho\rangle^{\otimes n}}{\left\|(\Pi^n_{\rho,\delta}\otimes I)\,|\phi_\rho\rangle^{\otimes n}\right\|}, \tag{5.10}$$

where we call the denominator $\sqrt{p} \geq \sqrt{1-\epsilon}$, with p being the probability of success. We want to show that this is $\approx |\phi_\rho\rangle^{\otimes n}$, which comes from gentle measurement: since our measurement succeeds most of the time, it cannot perturb the state that much (from pset 2). Here, "perturb the state too much" is easy to see in terms of fidelity, and fidelity and trace distance are related (by pset 1).

Explicitly, with $|\Phi\rangle = |\phi_\rho\rangle^{\otimes n}$ and $\Pi = \Pi^n_{\rho,\delta}\otimes I$:

$$|\langle\Phi|\psi\rangle|^2 = \frac{\langle\Phi|\,\Pi\,|\Phi\rangle^2}{p} = \frac{p^2}{p} = p \geq 1-\epsilon, \tag{5.11}$$

and since the output is $\sigma = p\,|\psi\rangle\langle\psi| + (1-p)\,\omega$ for some state ω, we get $F(|\Phi\rangle\langle\Phi|\,,\sigma) \geq p^2 \geq 1-2\epsilon$. □

This is achievability; we also need the converse, that we cannot do better. This will be left for the pset.

5.2 Relative entropy

Did you think we can escape compression? Nuh uh.

Back to classical compression: we have $X \sim p$, and we define

$$\text{surprise}(x) = \log\frac{1}{p(x)}, \qquad \text{so that} \qquad H(X) = \mathbb{E}[\text{surprise}(X)]. \tag{5.12}$$

Definition 5.2.1. *A code $E : [d] \to \{0,1\}^*$ is* prefix-free *if no codeword $E(x)$ is a prefix of another codeword $E(y)$, $y \neq x$.*

We can define a code with $|E(x)| = \lceil \text{surprise}(x)\rceil$ that is prefix-free. An example:

x	$p(x)$	$E(x)$
a	1/2	0
b	1/4	10
c	1/8	110
d	1/8	111

Prefix-freeness helps with decoding: for instance, to decode $1011010\ldots$, we eat digits until we get a valid codeword, and pop out the letter:

$$(10)(110)(10)\ldots \implies \text{bcb}\ldots. \tag{5.13}$$

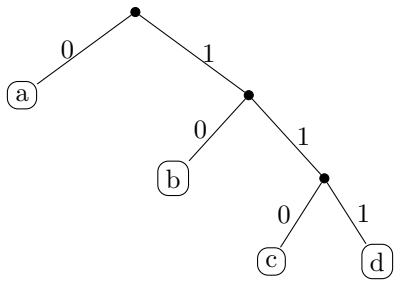


Figure 5.3: The code above as a binary tree. Codewords are exactly the leaves, which is why the code is prefix-free: no leaf lies on the path to another leaf. A letter with probability $2^{-\ell}$ sits at depth ℓ.

5.2.1 The wrong codebook

Different distributions of letters should have different encodings (say for French or English). But what if we use the wrong encoding, i.e. we encode French text (with some distribution p) using the English encoding (designed for q)? We want to find

$$\mathbb{E}_{x\sim p}|\mathrm{Enc}_q(x)| \approx \sum_x p(x)\log\frac{1}{q(x)} \geq \sum_x p(x)\log\frac{1}{p(x)}, \tag{5.14}$$

where the final inequality claims that the wrong encoding must do worse than the best encoding. We define this excess as follows.

Definition 5.2.2 (Relative entropy)**.** *For distributions p, q on $[d]$,*

$$D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)}. \tag{5.15}$$

Theorem 5.2.1 (Gibbs' inequality)**.** *$D(p\|q) \geq 0$, with equality if and only if $p = q$.*

One proof is operational: Shannon's coding theorem implies we cannot do better than the optimal code, so $D(p\|q)$ cannot be negative. Getting rid of the hand-waviness makes this a bit long (albeit good for intuition), so here is another proof.

Proof. First, note that $1 + z \leq e^z$ (standard Taylor inequality plus positive second derivative). Letting $y = e^z$, we can rewrite this as $\ln y \leq y - 1$, or equivalently $\ln(1/y) \geq 1 - y$. Then (ignoring 0 issues for now)

$$(\ln 2)\, D(p\|q) = \sum_x p(x)\ln\frac{p(x)}{q(x)} \geq \sum_x p(x)\left(1 - \frac{q(x)}{p(x)}\right) = \sum_x \big(p(x) - q(x)\big) = 0 \tag{5.16}$$

One nice thing about this proof is that it gives the equality conditions, namely when the ln bound is tight. That holds when $y = 1$, which implies $q(x) = p(x)$ for all x. Thus $D(p\|q) = 0 \iff p = q$. □

One note: what if $p(x) = 0$? As with entropy, we assume $0\log 0 = 0$. If $q(x) = 0$ but $p(x)$ is not, then we get claims like

$$D\left(\begin{pmatrix}0.5\\0.5\end{pmatrix}\middle\|\begin{pmatrix}1\\0\end{pmatrix}\right) = \infty. \tag{5.17}$$

A way to intuit this: if our codebook lacks a letter (where q is the codebook), the cost of encoding messages that require that letter is infinite.

D is not a distance: in general $D(p\|q) \neq D(q\|p)$, and there is no triangle inequality. But it is always nonnegative and zero only when $p = q$.

Corollary 5.2.1. $H(p) \leq \log d$.

Proof. This follows from the converse of the noiseless coding theorem, but we can argue using relative entropy as well. Let $u = (1/d, \dots, 1/d)$ be the uniform distribution; then

$$D(p\|u) = \sum_x p(x) \log\big(d\, p(x)\big) = -H(p) + \log d \geq 0, \tag{5.18}$$

giving $H(p) \leq \log d$. □

Corollary 5.2.2. $I(X;Y) \geq 0$.

Proof. The mutual information is a relative entropy:

$$D(p_{XY}\|p_X \otimes p_Y) = \sum_{x,y} p(x,y)\big(\log p(x,y) - \log p_X(x) - \log p_Y(y)\big) \tag{5.19}$$

$$= -H(XY) + H(X) + H(Y) = I(X;Y), \tag{5.20}$$

which is nonnegative by Theorem 5.2.1. □

This helps us intuitively understand mutual information as a measure of correlation: it measures how far the joint distribution is from the product of its marginals, and it vanishes exactly when X and Y are independent.

Alternatively, we can see $I(X;Y) \geq 0$ from concavity of entropy, using the definition $I(X;Y) = H(Y) - H(Y|X)$. Suppose $p(x,y) = \pi_x p_x(y)$, where π is a distribution of x and p_x is a distribution of y (given x). Then, we know

$$H(Y|X) = \sum_x \pi_x H(p_x), \qquad H(Y) = H\left(\sum_x \pi_x p_x\right), \tag{5.21}$$

and the entropy of the average is higher than the average of the entropy (think of the frowny face entropy of $H(p, 1-p)$, in Figure 5.4), so $I(X;Y) \geq 0$.

5.3 Hypothesis testing

Say x is drawn from either p or q. For trace distance, for instance, our calculations assumed equal priors. But now let us say the priors are uneven (or the two mistakes have different costs).

Definition 5.3.1. *For a test that guesses p or q from a sample x, let*

$$\alpha = \Pr[\textit{guess } q \mid x \sim p], \qquad \beta = \Pr[\textit{guess } p \mid x \sim q]. \tag{5.22}$$

The first is the type I error *and the second is the* type II error.

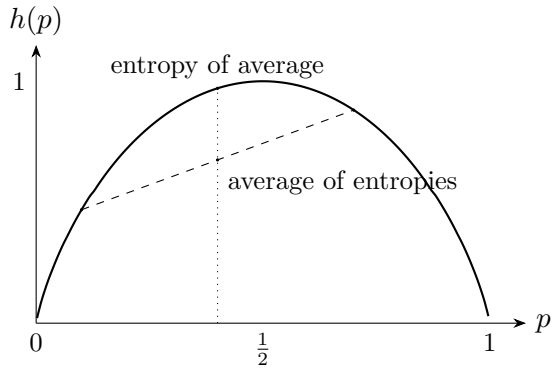


Figure 5.4: The "frowny face" binary entropy $h(p) = H(p, 1-p)$. Mixing $p = 0.1$ and $p = 0.7$ with equal weights: the chord (dashed) lies below the curve, so the entropy of the mixture exceeds the average of the entropies.

We might have a symmetric cost, where

$$\min \frac{\alpha + \beta}{2} = \frac{1}{2}\Big(1 - \frac{1}{2}\|p - q\|_1\Big), \tag{5.23}$$

or a Bayesian cost with prior π on p, where

$$\min\ \pi\alpha + (1-\pi)\beta = \frac{1}{2}\Big(1 - \|\pi p - (1-\pi)q\|_1\Big). \tag{5.24}$$

We will be interested in something asymmetric. Let

$$\beta_\epsilon = \min\{\beta : \alpha \le \epsilon\}, \tag{5.25}$$

and we care about $\beta_\epsilon^n \coloneqq (\beta_\epsilon \text{ for } p^{\otimes n} \text{ vs. } q^{\otimes n})$. Why ask this? Because it has a nice answer.

Theorem 5.3.1 (Chernoff–Stein lemma)**.** *For any $\epsilon \in (0,1)$,*

$$\lim_{n\to\infty} -\frac{1}{n}\log \beta_\epsilon^n = D(p\|q). \tag{5.26}$$

Informally, $\beta_\epsilon^n \approx 2^{-nD(p\|q)}$ as $n \to \infty$.

1. $p = q$**.** If $p = q$, then $\beta_\epsilon^n = 1 - \epsilon$, as the sample gives no information (the relative entropy is 0, so there is no exponential decay). If $p \neq q$, we claim $\beta_\epsilon^n \sim 2^{-nR}$ with $R = D(p\|q) > 0$.

2. $q = u$ **uniform.** If $q = u$, we have $D(p\|u) = \log d - H(p)$. After drawing x^n from either p or u: if $x^n \in T_{p,\delta}^n$ (the typical set), we guess p; otherwise we guess u. The type I error is $p^n\big(\overline{T}_{p,\delta}^n\big) \le \epsilon$ for large n. For the other direction,

$$\beta = u^n\big(T_{p,\delta}^n\big) = \frac{|T_{p,\delta}^n|}{d^n} \le \exp\big(n\ln 2\,(H(p) + \delta - \log d)\big) = \exp\big(-n\ln 2\,(D(p\|u) - \delta)\big). \tag{5.27}$$

3. $D(p\|q) = \infty$**.** Say $D(p\|q) = \infty$, meaning there is some x where $p(x) > 0$ and $q(x) = 0$. We can also state this as $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$ (the support is the set of points where the distribution is nonzero). What we do is if any sample lies outside $\mathrm{supp}(q)$, guess p; otherwise guess q. Then $\beta = 0$, since samples from q never lie outside $\mathrm{supp}(q)$, and

$$\alpha = p\big(\mathrm{supp}(q)\big)^n \to 0, \tag{5.28}$$

since $p(\mathrm{supp}(q)) < 1$. So for large n we get $\beta_\epsilon^n = 0$, consistent with an infinite exponent.

8.372 Quantum Information Science III **Fall 2026**

Lecture 6: September 29, 2026

Scribe: Aarush Vailaya *Quantum relative entropy*

6.1 Hypothesis testing

Recall the setup from last lecture. Given p, q, let β_ϵ^n be the lowest probability of guessing p^n given q^n for any test that guesses q^n given p^n with probability $\leq \epsilon$. The Chernoff–Stein theorem states $\beta_\epsilon^n \sim \exp(-nD(p\|q))$.

Proof sketch of achievability. We use a likelihood ratio test (LRT). Given $x^n \sim p^n$ or q^n, we want a binary test that given x^n outputs p or q. We will start with a real valued test

$$W(x^n) = \log\left[\frac{p^n(x^n)}{q^n(x^n)}\right] = \sum_{i=1}^n \log\frac{p(x_i)}{q(x_i)}, \tag{6.29}$$

called the likelihood ratio (note each x_i is an IID random variable). When we want to balance the two errors (for instance when calculating trace distance with the best measurement), we would want to guess p when $W \geq 0$ and q otherwise. But we are doing asymmetrical hypothesis testing, so we need to pick a threshold T for W. Define

$$A = \{x^n : W(x^n) \geq T\}, \tag{6.30}$$

where our binary test outputs p^n for all $x^n \in A$ and q^n for all $x^n \notin A$. Note

$$\mathbb{E}_{x^n\sim p^n}[W] = nD(p\|q), \qquad \mathbb{E}_{x^n\sim q^n}[W] = -nD(q\|p). \tag{6.31}$$

Let us set $T = n(D(p\|q) - \delta)$; we need to find the probability that x^n lands in A when we sample from p^n and sample from q^n. Note

$$x^n \in A \iff p^n(x^n) \geq 2^{n(D(p\|q)-\delta)} q^n(x^n) \tag{6.32}$$

(which comes from algebraic manipulations on the definition of A), from which we can upper bound $q^n(A)$ as

$$q^n(A) \leq 2^{-n(D(p\|q)-\delta)} p^n(A) \leq 2^{-n(D(p\|q)-\delta)}, \tag{6.33}$$

which is indeed exponential. We also need to show the other type of error (guessing q when we have p) is at most ϵ, meaning we need to show

$$1 - p^n(A) \leq \epsilon \tag{6.34}$$

for large enough n. This is true due to the weak law of large numbers. The weak law says the average of I.I.D. variables approaches a point. With very high probability, W/n will be very close to its expectation, $D(p\|q)$. In particular,

$$\Pr_{x^n\sim p^n}[W/n \geq D(p\|q) - \delta] \to 1 \text{ as } n \to \infty. \tag{6.35}$$

□

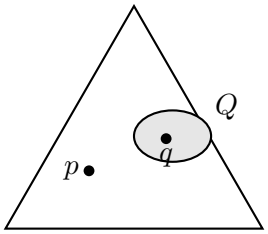


Figure 6.5: Multiple hypothesis testing on the probability simplex: the null hypothesis p is a single point, and the alternative is any q in the set Q.

The above is an achievability proof of Chernoff–Stein, and this can be extended to multiple hypothesis testing. In particular, imagine the probability simplex in Figure 6.5, where $x^n \sim p^n$ or $x^n \sim q^n$ for $q \in Q$; our test has to work against all of these hypotheses simultaneously.

The converse will be proven as part of the quantum variant of Chernoff–Stein.

6.2 Quantum relative entropy

Quantum relative entropy is defined as

$$D(\rho\|\sigma) = \operatorname{tr}[\rho(\log\rho - \log\sigma)], \tag{6.36}$$

which is ∞ if and only if $\operatorname{supp}\rho \not\subseteq \operatorname{supp}\sigma$ (where $\operatorname{supp}\rho$ is the subspace on which it is invertible, i.e. the span of the eigenvectors of ρ with nonzero eigenvalues; since ρ is Hermitian, this is also its range). Intuitively, we get infinity when we have a zero eigenvalue in σ paired with a non-zero eigenvalue in ρ.

An example of this is $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$. If $|\psi\rangle = |\phi\rangle$ then $D = 0$, and infinity otherwise (since the one-dimensional support of ρ is not contained in the support of σ).

We now have quantum Stein's lemma.

Theorem 6.2.1 (Quantum Stein's lemma)**.** *Define*

$$\beta_\epsilon^n = \min\left\{\operatorname{tr} M\sigma^{\otimes n} : 0 \le M \le I,\ \operatorname{tr} M\rho^{\otimes n} \ge 1-\epsilon\right\}, \tag{6.37}$$

over measurement matrices M, then

$$\lim_{n\to\infty} \frac{-1}{n}\log\beta_\epsilon^n = D(\rho\|\sigma). \tag{6.38}$$

For another example (with ρ, σ as above) introduce $A = |\phi^\perp\rangle\langle\phi^\perp|$, $B = I - A$, where $\operatorname{tr} A\rho > 0$. Consider $M = I - B^{\otimes n}$, which is equivalent to doing the A/B measurement n times; if we ever get A, we know we do not have $|\phi\rangle$. Then,

$$\operatorname{tr} M\sigma^{\otimes n} = 0, \qquad \operatorname{tr} M\rho^{\otimes n} = 1 - (1 - \operatorname{tr} A\rho)^n \to 1. \tag{6.39}$$

So M accepts $\rho^{\otimes n}$ with probability tending to 1 and never accepts $\sigma^{\otimes n}$. In the setting of Theorem 6.2.1, $\beta_\epsilon^n = 0$ for large enough n, so the error exponent is infinite, matching $D(\rho\|\sigma) = \infty$.

It also turns out $D(\rho\|\sigma) \ge 0$, which will be proven later. Intuitively, we may want to work in a basis where ρ is diagonal, which almost reduces $D(\rho\|\sigma)$ to the classical case until we hit the $\log(\sigma)$, after which we need to put in more work to show the inequality doesn't break. Assuming it for now, it implies $S(A) \le \log d_A$ (take $\sigma = I/d_A$: $D(\rho_A\|I/d_A) = \log d_A - S(A)$).

It also proves $I(A;B)_\rho = D(\rho_{AB}\|\rho_A \otimes \rho_B) \geq 0$ in the quantum case. The above statement is equivalent to $S(A) \geq S(A \mid B)$. Although conditional entropy has a nice classical interpretation, the quantum conditional entropy has no such interpretation. Nevertheless, the inequality holds.

Now, we want to relate $D(\rho\|\sigma)$ to trace distance and fidelity. One useful relation is Pinsker's inequality:

$$D(\rho\|\sigma) \geq \frac{1}{2\ln 2}\|\rho - \sigma\|_1^2, \tag{6.40}$$

where we proved a version of this on the pset. Intuitively, we can think of this operationally; the trace distance says there is a 2-valued POVM (test) that has a bias at least the trace distance. In particular, denoting the trace distance as $\epsilon = \frac{1}{2}\|\rho - \sigma\|_1$, then there exists a test $\{M, I - M\}$ and some p such that the probability of guessing ρ given it is ρ is at least $p + \epsilon$ and given it is σ at most p, i.e. $\operatorname{tr} M\rho \geq p + \epsilon$ and $\operatorname{tr} M\sigma \leq p$. This test can be used to distinguish n copies of ρ from n copies of σ, which when set up properly can give an exponentially decaying error, which bounds $D(\rho\|\sigma)$.

Concretely, measure each copy with $\{M, I - M\}$. This turns the problem into distinguishing n coin flips with heads probability $a = \operatorname{tr} M\rho \geq p + \epsilon$ from n flips with heads probability $b = \operatorname{tr} M\sigma \leq p$. By the classical Chernoff–Stein lemma this test achieves error exponent $D(a\|b)$ (the relative entropy of the two coins), and by Theorem 6.2.1 no test beats $D(\rho\|\sigma)$, so

$$D(\rho\|\sigma) \geq D(a\|b) \geq \frac{2}{\ln 2}(a-b)^2 \geq \frac{2}{\ln 2}\epsilon^2 = \frac{1}{2\ln 2}\|\rho - \sigma\|_1^2, \tag{6.41}$$

using the two-outcome case of the classical Pinsker inequality.

Why is the trace distance squared? (This was asked in class.) If we want to distinguish a coin with heads probability 1 from one with $1 - \epsilon$, we need about $1/\epsilon$ flips. But to distinguish $1/2$ from $1/2 - \epsilon$, we need about $1/\epsilon^2$ flips; and we need to be pessimistic since we don't know which regime (p from the trace distance fact) we are in. Since distinguishing at rate D takes about $1/D$ samples, the worst case $D \approx \epsilon^2$ is what gives the square.

6.3 Proof of quantum Stein's lemma

We want M such that $\operatorname{tr} M\rho^{\otimes n} \geq \alpha$ and $\operatorname{tr} M\sigma^{\otimes n} \leq 2^{-nR}$ with $R \approx D(\rho\|\sigma)$. We can define a quantum LRT (which is actually optimal) which is

$$M = \left[\frac{\rho^{\otimes n}}{\alpha} - 2^{nR}\sigma^{\otimes n} \geq 0\right], \tag{6.42}$$

where the notation means we project onto the eigenvectors of $\rho^{\otimes n}/\alpha - 2^{nR}\sigma^{\otimes n}$ with nonnegative eigenvalues; notation-wise, think of it as an indicator function on eigenvalues.

We will not analyze this test directly. Instead we follow a proof due to Bjelaković and Siegmund-Schultze. Diagonalize

$$\rho = \sum_x r_x \left|\alpha_x\right\rangle\!\left\langle\alpha_x\right|, \qquad \sigma = \sum_x s_x \left|\beta_x\right\rangle\!\left\langle\beta_x\right|, \tag{6.43}$$

and define the typical projector

$$\Pi^n_{\rho\|\sigma,\delta} = \sum_{\substack{x^n \text{ s.t.} \\ \left|\frac{1}{n}\sum_i \log s_{x_i} - \operatorname{tr}\rho\log\sigma\right| \leq \delta}} \beta_{x^n}, \tag{6.44}$$

(here β_{x^n} denotes the projector $|\beta_{x^n}\rangle\langle\beta_{x^n}|$, with $|\beta_{x^n}\rangle = |\beta_{x_1}\rangle \otimes \cdots \otimes |\beta_{x_n}\rangle$, and similarly for α_{x^n}) which we compare with

$$\Pi^n_{\rho,\delta} = \sum_{\substack{x^n \text{ s.t.} \\ \left|-\frac{1}{n}\sum_i \log r_{x_i} - S(\rho)\right| \le \delta}} \alpha_{x^n}. \tag{6.45}$$

Both are typical sets in the usual sense: Π_ρ keeps strings whose average $-\log r_{x_i}$ is close to $S(\rho)$, and $\Pi_{\rho\|\sigma}$ keeps strings whose average $\log s_{x_i}$ is close to $\langle \log\sigma\rangle_\rho = \operatorname{tr}\rho\log\sigma$.

Both are typical for $\rho^{\otimes n}$: by the weak law of large numbers,

$$\operatorname{tr}\rho^{\otimes n}\Pi^n_{\rho\|\sigma,\delta} \ge 1-\epsilon, \qquad \operatorname{tr}\rho^{\otimes n}\Pi^n_{\rho,\delta} \ge 1-\epsilon, \tag{6.46}$$

where $\Pi^n_{\rho,\delta}$ commutes with $\rho^{\otimes n}$ and $\Pi^n_{\rho\|\sigma,\delta}$ commutes with $\sigma^{\otimes n}$, since each projector is diagonal in the eigenbasis of that state. From here on we write Π_ρ and $\Pi_{\rho\|\sigma}$ for $\Pi^n_{\rho,\delta}$ and $\Pi^n_{\rho\|\sigma,\delta}$. Note we have

$$\sigma^{\otimes n} = \sum_{x^n}\Big(\prod_i s_{x_i}\Big)\,|\beta_{x^n}\rangle\langle\beta_{x^n}|\,, \tag{6.47}$$

and $\Pi_{\rho\|\sigma}$ is diagonal in this same basis, so sandwiching $\sigma^{\otimes n}$ with it simply deletes the atypical eigenvectors:

$$\Pi_{\rho\|\sigma}\sigma^{\otimes n}\Pi_{\rho\|\sigma} = \sum_{x^n \text{ typical}}\Big(\prod_i s_{x_i}\Big)\,|\beta_{x^n}\rangle\langle\beta_{x^n}|\,. \tag{6.48}$$

(Since the two commute, this also equals $\Pi_{\rho\|\sigma}\sigma^{\otimes n}$. We write it as a sandwich because it is the part of $\sigma^{\otimes n}$ that passes the measurement $\Pi_{\rho\|\sigma}$, and because it is Hermitian, so operator inequalities make sense.) For a typical string, each surviving eigenvalue $\prod_i s_{x_i} = 2^{\sum_i \log s_{x_i}}$ lies between $2^{n(\operatorname{tr}\rho\log\sigma-\delta)}$ and $2^{n(\operatorname{tr}\rho\log\sigma+\delta)}$. The zero eigenvalues of $\sigma^{\otimes n}$ are never typical, since they have $\log s_{x_i} = -\infty$ for some i (we assume $D(\rho\|\sigma) < \infty$). So we get

$$2^{n(\operatorname{tr}\rho\log\sigma-\delta)}\Pi_{\rho\|\sigma} \le \Pi_{\rho\|\sigma}\sigma^{\otimes n}\Pi_{\rho\|\sigma} \le 2^{n(\operatorname{tr}\rho\log\sigma+\delta)}\Pi_{\rho\|\sigma}. \tag{6.49}$$

In the same way, the typical projector of ρ controls the eigenvalues of $\rho^{\otimes n}$:

$$2^{-n(S(\rho)+\delta)}\Pi_\rho \le \Pi_\rho\rho^{\otimes n}\Pi_\rho \le 2^{-n(S(\rho)-\delta)}\Pi_\rho. \tag{6.50}$$

So Π_ρ restricts the eigenvalues of $\rho^{\otimes n}$ to typical ones and $\Pi_{\rho\|\sigma}$ restricts the eigenvalues of $\sigma^{\otimes n}$, and both keep most of the weight of $\rho^{\otimes n}$.

Now, let us prove achievability using a nifty quantum trick. Classically, we might say we want a string typical with respect to both conditions (typical for Π_ρ and for $\Pi_{\rho\|\sigma}$, i.e. accept with $\Pi_\rho\Pi_{\rho\|\sigma}$, which is a projector when the two commute). If the two projectors don't commute, $\Pi_\rho\Pi_{\rho\|\sigma}$ is not even a projector; but what we can do is measure one projector, and if that succeeds, measure the second one. The order we do them in matters (they don't commute), but if both succeed with overwhelming probability, it won't matter much (gentle measurement). In particular, let

$$M = \Pi_{\rho\|\sigma}\Pi_\rho\Pi_{\rho\|\sigma}, \tag{6.51}$$

meaning

$$\operatorname{tr} M\rho^{\otimes n} = \operatorname{tr}\Pi_\rho(\Pi_{\rho\|\sigma}\rho^{\otimes n}\Pi_{\rho\|\sigma}). \tag{6.52}$$

This is *sequential measurement*. We know both projectors on their own succeed with overwhelming probability, and the combination does as well, since

$$\Pi_{\rho\|\sigma}\rho^{\otimes n}\Pi_{\rho\|\sigma} \underset{\sqrt{2\epsilon}}{\approx} \rho^{\otimes n}, \tag{6.53}$$

so

$$\operatorname{tr} M\rho^{\otimes n} = \operatorname{tr}\Pi_\rho\rho^{\otimes n} - \operatorname{tr}\Pi_\rho\big(\rho^{\otimes n} - \Pi_{\rho\|\sigma}\rho^{\otimes n}\Pi_{\rho\|\sigma}\big) \tag{6.54}$$

$$\geq 1 - \epsilon - \sqrt{2\epsilon}. \tag{6.55}$$

Here $\operatorname{tr}\Pi_\rho\rho^{\otimes n} \geq 1-\epsilon$, and the second term is at most $\sqrt{2\epsilon}$ by gentle measurement (pset 2), since $\operatorname{tr}\Pi_{\rho\|\sigma}\rho^{\otimes n} \geq 1-\epsilon$.

For the type II error, we have (using (6.49), $\Pi_{\rho\|\sigma} \leq I$, and $\operatorname{tr}\Pi_\rho \leq 2^{n(S(\rho)+\delta)}$)

$$\operatorname{tr} M\sigma^{\otimes n} = \operatorname{tr}\Pi_\rho\left(\Pi_{\rho\|\sigma}\sigma^{\otimes n}\Pi_{\rho\|\sigma}\right) \tag{6.56}$$

$$\leq 2^{n(\operatorname{tr}\rho\log\sigma+\delta)}\operatorname{tr}\Pi_\rho \tag{6.57}$$

$$\leq \exp\big(n\ln 2\,(\operatorname{tr}\rho\log\sigma + \delta) + n\ln 2\,(S(\rho)+\delta)\big) \tag{6.58}$$

$$= \exp\big(-n\ln 2\,(D(\rho\|\sigma) - 2\delta)\big). \tag{6.59}$$

This shows achievability; now we show the converse. Suppose a test M has $\operatorname{tr} M\rho^{\otimes n} \geq \alpha$. We want to argue $\operatorname{tr} M\sigma^{\otimes n} \gtrsim 2^{-nD(\rho\|\sigma)}$. We can use the same inequality from earlier but use the lower bound instead of the upper bound in (6.49). We get

$$\sigma^{\otimes n} \geq 2^{n(\operatorname{tr}\rho\log\sigma-\delta)}\Pi_{\rho\|\sigma}. \tag{6.60}$$

This holds because $\Pi_{\rho\|\sigma}$ commutes with $\sigma^{\otimes n}$, so $\sigma^{\otimes n} = \Pi_{\rho\|\sigma}\sigma^{\otimes n}\Pi_{\rho\|\sigma} + (I-\Pi_{\rho\|\sigma})\sigma^{\otimes n}(I-\Pi_{\rho\|\sigma})$. The second term is positive semidefinite (it contains any zero eigenspace of $\sigma^{\otimes n}$), so dropping it only makes the right side smaller, and then we apply the lower bound in (6.49). Then,

$$\operatorname{tr} M\sigma^{\otimes n} \geq 2^{n(\operatorname{tr}\rho\log\sigma-\delta)}\operatorname{tr}(M\Pi_{\rho\|\sigma}), \tag{6.61}$$

we want to show this is not too small. We have

$$\operatorname{tr} M\Pi_{\rho\|\sigma} = \operatorname{tr}\Pi_{\rho\|\sigma}M\Pi_{\rho\|\sigma} \geq \operatorname{tr}\Pi_\rho\Pi_{\rho\|\sigma}M\Pi_{\rho\|\sigma} \tag{6.62}$$

(since $\Pi_{\rho\|\sigma}M\Pi_{\rho\|\sigma} \geq 0$ and $\Pi_\rho \leq I$). Side calculation:

$$\rho^{\otimes n} = \Pi_\rho\rho^{\otimes n} + (I-\Pi_\rho)\rho^{\otimes n}, \tag{6.63}$$

and the typical projector satisfies

$$\Pi_\rho \geq \left(\rho^{\otimes n} - (I-\Pi_\rho)\rho^{\otimes n}\right)2^{n(S(\rho)-\delta)}. \tag{6.64}$$

To see this: $\Pi_\rho\rho^{\otimes n} = \Pi_\rho\rho^{\otimes n}\Pi_\rho \leq 2^{-n(S(\rho)-\delta)}\Pi_\rho$ by (6.50), and $\Pi_\rho\rho^{\otimes n} = \rho^{\otimes n} - (I-\Pi_\rho)\rho^{\otimes n}$ by the line above. So going back to the main thing, we can bound

$$\operatorname{tr} M\Pi_{\rho\|\sigma} \geq \operatorname{tr}\big((\rho^{\otimes n} - (I-\Pi_\rho)\rho^{\otimes n})\Pi_{\rho\|\sigma}M\Pi_{\rho\|\sigma}\big)\,2^{n(S(\rho)-\delta)}. \tag{6.65}$$

The first term is a sequential measurement on $\rho^{\otimes n}$: first $\Pi_{\rho\|\sigma}$, which barely disturbs the state, then M, which accepts $\rho^{\otimes n}$ with probability at least α. The second term is small because Π_ρ captures most of $\rho^{\otimes n}$. So the above is

$$\geq (\alpha - \sqrt{2\epsilon} - \epsilon)2^{n(S(\rho)-\delta)}. \tag{6.66}$$

Putting it all together, we get

$$\operatorname{tr} M\sigma^{\otimes n} \geq (\alpha - \sqrt{2\epsilon} - \epsilon)\exp\big(-n\ln 2\,(D(\rho\|\sigma) + 2\delta)\big). \tag{6.67}$$

Thus, if M accepts ρ with probability $\geq \alpha$, it accepts σ with a probability that is at least about $2^{-nD(\rho\|\sigma)}$, so no test has a better error exponent than $D(\rho\|\sigma)$. This is not the first proof; the original proof relied on strong subadditivity (Section 6.4), which itself was proven using difficult algebra.

6.4 Monotonicity of relative entropy

A simple corollary is monotonicity under any channel $\mathcal{E}$:

$$D(\rho\|\sigma) \geq D(\mathcal{E}(\rho)\|\mathcal{E}(\sigma)), \tag{6.68}$$

since achievability means there is a measurement M on $\mathcal{E}(\rho)^{\otimes n}$, $\mathcal{E}(\sigma)^{\otimes n}$ achieving exponent $D(\mathcal{E}(\rho)\|\mathcal{E}(\sigma))$, so we can simply apply $\mathcal{E}^{\otimes n}$ followed by M on $\rho^{\otimes n}$, $\sigma^{\otimes n}$. This test has the same errors, and by the converse no test on $\rho^{\otimes n}$ vs. $\sigma^{\otimes n}$ beats $D(\rho\|\sigma)$; so we need both halves of Theorem 6.2.1. This gives us monotonicity of the relative entropy.

Suppose that $\rho = \rho_{ABC}$ and $\sigma = \rho_A \otimes \rho_{BC}$ and $\mathcal{E} = \mathrm{tr}_C$. Note

$$D(\rho\|\sigma) = I(A;BC)_\rho \geq I(A;B) = D(\mathcal{E}(\rho)\|\mathcal{E}(\sigma)), \tag{6.69}$$

meaning the amount that A knows of BC is at least as much as it knows about B. This is called strong subadditivity, and this difference is sometimes called the conditional mutual information,

$$I(A;C \mid B) = I(A;BC) - I(A;B) \geq 0. \tag{6.70}$$

Monotonicity also gives the data processing inequality: suppose we want to send a classical message through a quantum channel. We have classical X which we turn to quantum state Q and then output Y; then, $I(X;Y) \leq I(X;Q)$. Isometries don't change entropies; only throwing away subsystems does (every channel is an isometry followed by throwing away a subsystem).

How does the above give data processing? Imagine the channel that measures Q and gives us Y. Applying monotonicity with this channel acting on Q, to ρ_{XQ} vs. $\rho_X \otimes \rho_Q$, gives

$$I(X;Y) = D(\rho_{XY}\|\rho_X \otimes \rho_Y) \leq D(\rho_{XQ}\|\rho_X \otimes \rho_Q) = I(X;Q). \tag{6.71}$$

We will return to this in a future lecture.

8.372 Quantum Information Science III **Fall 2026**

Lecture 7: October 1, 2026

Scribe: Aarush Vailaya *Quantum sensing and Fisher information*

Guest lecturer: Soonwon Choi.

7.1 Quantum sensing in practice

For more background, see C. L. Degen, F. Reinhard, and P. Cappellaro, *Quantum sensing*, Rev. Mod. Phys. **89**, 035002 (2017).

Suppose we have a qubit exposed to some external field (say, a magnetic or electric field), described by the Hamiltonian

$$H = -\frac{\Delta}{2}\sigma_z, \tag{7.72}$$

and we want to estimate the strength of the field, Δ. This is a well-understood problem with an information-theoretically optimal approach, known as Ramsey spectroscopy, which is what we will go over.

First, prepare $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) = |+\rangle$. Then, evolve for a time T under $U(T) = e^{-iHT}$. Next, perform a measurement, for example in the σ_y basis (as we will see, any measurement axis in the equatorial plane works equally well). Finally, repeat this M times and use the outcomes to estimate Δ. Note that we haven't yet fixed the time T or the number of repetitions M; these are parameters we will optimize.

Consider the Bloch sphere (Figure 7.6), with $|0\rangle$ and $|1\rangle$ at the poles and the qubit initially on the equator. After the unitary, it has rotated about the z axis by an angle ΔT. Thus, our measurement in the σ_y basis has a certain probability of each outcome; for instance, if the rotation is precisely $\pi/2$, we are guaranteed to measure the positive eigenvalue of σ_y. In particular,[1]

$$\Pr(+_y) = \frac{1}{2}(1 + \sin(\Delta T)), \qquad \Pr(-_y) = \frac{1}{2}(1 - \sin(\Delta T)). \tag{7.73}$$

If we plot $\Pr(+_y)$ against ΔT, we get an oscillating curve, the Ramsey fringe, starting at $1/2$ (Figure 7.7). Assuming we have done this measurement before, we have some estimate Δ_0, and we write $\Delta = \Delta_0 + \delta\Delta$, where $\delta\Delta$ is unknown but small ($|\delta\Delta|\, T < \pi/2$). We cannot change Δ, but we can shift the fringe by a known phase β, using a known bias field or by rotating the initial state or the measurement axis about z. The accumulated phase is then $(\Delta_0 T + \beta) + \delta\Delta\, T$: the known part moves freely along the fringe, while $\delta\Delta$ is unchanged. From now on we absorb β into Δ_0. For instance, we can pick Δ_0 so that $\Delta_0 T$ sits on the steepest part of the fringe, and because of the uncertainty the true value may be slightly higher or lower. But what is the optimal Δ_0 to pick?

The signal is the change in the outcome probability P caused by $\delta\Delta$:[2]

$$\delta p = \left.\frac{\partial P}{\partial \Delta}\right|_{\Delta=\Delta_0} \cdot\, \delta\Delta = 2\cos(\Delta T/2)\sin(\Delta T/2) \cdot \frac{T}{2}\,\delta\Delta. \tag{7.74}$$

[1] With the sign of H in (7.72), the Bloch vector actually rotates clockwise about z, so strictly $\Pr(\pm_y) = \frac{1}{2}(1 \mp \sin(\Delta T))$. The overall sign of the signal plays no role in what follows.

[2] On the board, $P = \sin^2(\Delta T/2)$, which is (7.73) shifted by $\pi/2$ (as from measuring σ_x instead); this makes no difference.

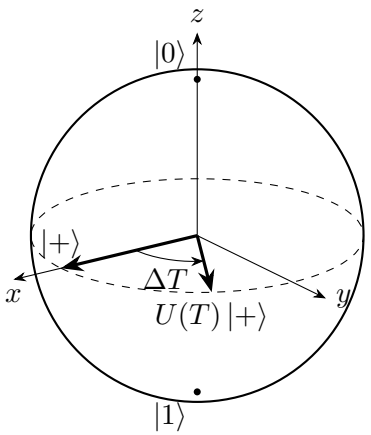


Figure 7.6: Ramsey spectroscopy on the Bloch sphere. The qubit starts at $|+\rangle$ on the equator and precesses about the z axis by an angle ΔT; measuring σ_y then reads out $\sin(\Delta T)$.

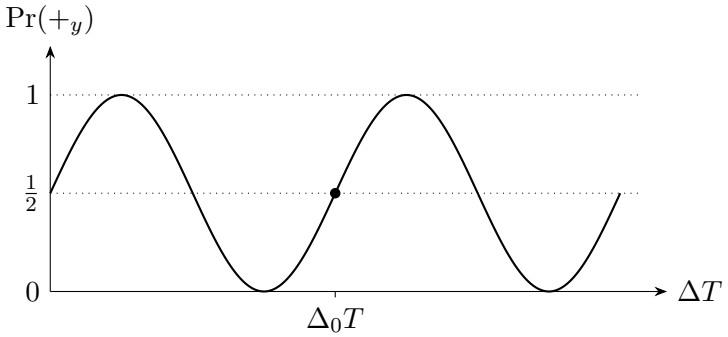


Figure 7.7: The Ramsey fringe (7.73). Here we choose Δ_0 so that $\Delta_0 T$ sits at the steepest point of the fringe (dot); a small $\delta\Delta$ then shifts $\Pr(+_y)$ linearly.

The repetitions give outcomes $Y_1 = \pm 1, Y_2 = \pm 1, \ldots$, where we call $+1$ the outcome with probability P; based on M repetitions we can estimate $\hat{p} = \#(+1)/M$. We cannot read off the value of the field; we can only read off the discrete outcomes of the quantum mechanical measurement. Each repetition is a coin flip with bias P, so

$$\mathrm{Var}(\hat{p}) = \frac{1}{M}\mathrm{Var}_1(\hat{p}) = \frac{1}{M}P(1-P) = \frac{1}{M}\sin^2(\Delta T/2)\cos^2(\Delta T/2), \tag{7.75}$$

$$\mathrm{Std}(\hat{p}) = \frac{1}{\sqrt{M}}|\sin(\Delta T/2)\cos(\Delta T/2)|, \tag{7.76}$$

where Var_1 is the variance of a single repetition. Given M, we care about the signal-to-noise ratio (SNR),

$$\mathrm{SNR} = \frac{|\delta p|}{\mathrm{Std}(\hat{p})} = \frac{\left|2\cos(\Delta T/2)\sin(\Delta T/2)\cdot \frac{T}{2}\,\delta\Delta\right|}{\frac{1}{\sqrt{M}}|\sin(\Delta T/2)\cos(\Delta T/2)|} = \sqrt{M}\,T\,|\delta\Delta|. \tag{7.77}$$

We say the measurement works if $\mathrm{SNR} \geq 1$, which means

$$\delta\Delta_{\min} \approx \frac{1}{T}\cdot\frac{1}{\sqrt{M}}, \tag{7.78}$$

and notice that this does not depend on Δ_0. The intuition is that there is a tradeoff between signal and projection noise: the factor $|\sin(\Delta T/2)\cos(\Delta T/2)|$ appears in both and cancels. On flat regions of the fringe the signal is low, but so is the noise; on sloped regions the signal is high, but so is the noise.

In these problems, we have a fixed resource, the total time $\tau = T \cdot M$, so (substituting $M = \tau/T$) we need

$$\delta\Delta_{\min} \approx \frac{1}{\sqrt{T \cdot \tau}}. \tag{7.79}$$

T is a parameter we can choose in the protocol, so we want to make it as large as possible (but note $T \leq \tau$). Taking $T = \tau$ gives $\delta\Delta_{\min} \approx 1/\tau$, which is called the Heisenberg limit (here with respect to the total time τ). In practice, decoherence prevents this: the fringe decays as

$$\Pr(+_y) = \frac{1}{2}\left(1 + \sin(\Delta T)e^{-T/T_2}\right), \tag{7.80}$$

where T_2 is the coherence time of the qubit. Note that T is the evolution time we choose in each repetition, while T_2 is a fixed property of the qubit. In this scenario, repeating the math above gives $\delta\Delta_{\min} \approx e^{T/T_2}/\sqrt{T\tau}$, which is smallest for $T \approx T_2$ (the exact optimum is $T = T_2/2$, but we drop constant factors), giving

$$\delta\Delta_{\min} \approx \frac{1}{\sqrt{T_2 \cdot \tau}}. \tag{7.81}$$

If we use N particles in parallel, the number of repetitions M becomes N times larger, so

$$\delta\Delta_{\min} \approx \frac{1}{\sqrt{NT_2\tau}}, \tag{7.82}$$

which is called the standard quantum limit. The word "quantum" comes from the fact that our precision is limited by projection noise: even with a perfect apparatus, each outcome is random by the Born rule, and the average of N independent outcomes fluctuates by about $1/\sqrt{N}$.

But there is a better approach: with N particles, prepare the GHZ state $\frac{1}{\sqrt{2}}(|0\cdots0\rangle + |1\cdots1\rangle)$, which behaves like a single logical qubit. Under $H = -\frac{\Delta}{2}\sum_i \sigma_z^{(i)}$, the energy separation between $|0\cdots0\rangle$ and $|1\cdots1\rangle$ is $N\Delta$, so the relative phase accumulates N times faster, which gives

$$\delta\Delta_{\min} = \frac{1}{N} \cdot \frac{1}{\sqrt{T_{2,\mathrm{GHZ}}\,\tau}}. \tag{7.83}$$

For the GHZ state, we measure in the basis of even and odd GHZ states, $\frac{1}{\sqrt{2}}(|0\cdots0\rangle \pm |1\cdots1\rangle)$.

The protocols above are simplified. For example, we assumed that Δ_0 is already known well enough ($|\delta\Delta| < \pi/(2T)$) to sit on a steep part of the fringe; the range of fields a protocol can handle without such prior knowledge is called its dynamic range. One issue with a single GHZ state is that it decoheres faster: if each qubit dephases independently, then $T_{2,\mathrm{GHZ}} = T_2/N$, and so

$$\frac{1}{N} \cdot \frac{1}{\sqrt{T_{2,\mathrm{GHZ}}\,\tau}} = \frac{1}{N} \cdot \sqrt{\frac{N}{T_2\tau}} = \frac{1}{\sqrt{NT_2\tau}}, \tag{7.84}$$

which is no better than the standard quantum limit.

7.2 Fisher information and the Cramér–Rao bound

Suppose we have two distributions $P_0(x)$ and $P_1(x)$ with a 50-50 prior, and we get a sample x_*. The optimal guess is to check whether $P_0(x_*) \le P_1(x_*)$ and return whichever hypothesis gives the higher probability. With n samples we can do something similar, using $P_{0/1}(\vec{x}) = \prod_j P_{0/1}(x_j)$. Let us define the statistic

$$W_n(\vec{x}) = \log \frac{\prod_j P_1(x_j)}{\prod_j P_0(x_j)} = \sum_j \log \frac{P_1(x_j)}{P_0(x_j)}, \tag{7.85}$$

whose sign tells us the optimal guess (just like before). Since W_n is a sum of n i.i.d. terms, the law of large numbers and the central limit theorem tell us how it behaves for large n. Suppose we are actually drawing from P_1; then

$$\lim_{n\to\infty} \frac{1}{n} W_n(\vec{x}) = \sum_x P_1(x) \log \frac{P_1(x)}{P_0(x)} = D(P_1 \| P_0) \ge 0, \tag{7.86}$$

which gives an intuitive reason why the relative entropy should be nonnegative: we expect W_n to be positive since we are actually sampling from P_1.

Pictorially (Figure 7.8), as we draw more samples, W_n performs a random walk with average slope $D(P_1\|P_0)$. At any fixed n, the distribution of W_n is approximately Gaussian, and its variance grows linearly in n, so its width grows only like $\sqrt{n}$ while its mean grows like n.

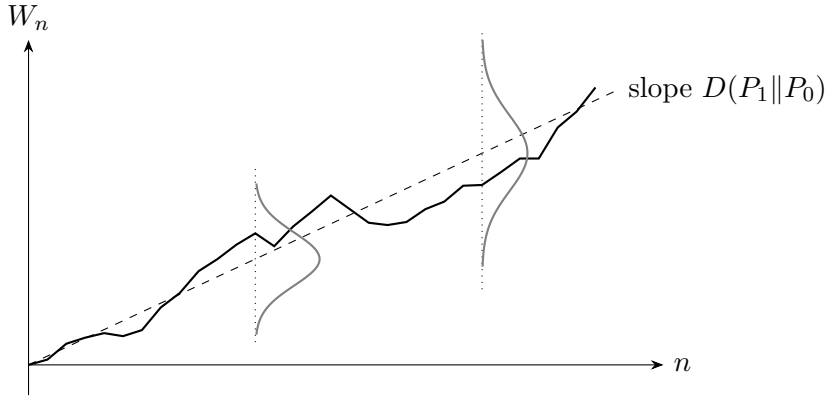


Figure 7.8: The log-likelihood ratio W_n as a random walk when sampling from P_1. Its mean grows like $n\,D(P_1\|P_0)$ (dashed), while at fixed n it is approximately Gaussian (gray) with width growing like $\sqrt{n}$.

Now say we have $P_0(x)$ and a family $P_\theta(x)$ depending smoothly on one parameter, where $\theta = \theta_0 + \delta\theta$; for simplicity we take $\theta_0 = 0$, so $\delta\theta = \theta$. Let $W_n = \sum_j \log \frac{P_\theta(x_j)}{P_0(x_j)}$, and sample from P_θ. Then

$$\overline{W_n} = nD(P_\theta\|P_0), \tag{7.87}$$

$$\mathrm{Var}(W_n) = n\,\mathrm{Var}(W_1) = n\left[-D(P_\theta\|P_0)^2 + \sum_x P_\theta(x)\left(\log\frac{P_\theta(x)}{P_0(x)}\right)^2\right]. \tag{7.88}$$

We are interested in the regime where θ is very small. So first, let us expand the mean around $\theta = 0$:

$$\overline{W}_n = nD(P_\theta\|P_0) = 0 + 0\,\delta\theta + n\underbrace{\partial_\theta^2 D(P_\theta\|P_0)\big|_{\theta=0}}_{=F}\cdot\frac{\delta\theta^2}{2} + O(\delta\theta^3). \tag{7.89}$$

The zeroth-order term vanishes because $D(P_0\|P_0) = 0$, and the first-order term vanishes because D cannot be negative, so $\theta = 0$ is a minimum. The second derivative F is the *Fisher information*. For the sake of time we skip the calculation, but expanding $\log \frac{P_\theta}{P_0} \approx \theta\, \partial_\theta \log P_\theta$ in the same way, we get

$$\mathrm{Var}(W_n) = n\theta^2 F + O(\theta^3), \tag{7.90}$$

where

$$F = \sum_x P_\theta(x) \left(\frac{\partial_\theta P_\theta(x)}{P_\theta(x)} \right)^2 \Big|_{\theta=\theta_0} = \sum_x \frac{(\partial_\theta P_\theta(x))^2}{P_\theta(x)} \Big|_{\theta=\theta_0}. \tag{7.91}$$

So, consider the normalized statistic

$$Z_n = \frac{W_n}{\sqrt{\mathrm{Var}(W_n)}}. \tag{7.92}$$

By the law of large numbers and the central limit theorem, it is approximately Gaussian with standard deviation 1 and mean

$$\overline{Z_n} = \frac{nF\theta^2/2}{\sqrt{nF\theta^2}} = \frac{\sqrt{nF\theta^2}}{2}. \tag{7.93}$$

We guess correctly when $Z_n > 0$ and wrongly when $Z_n < 0$ (Figure 7.9). To keep the error probability below some ϵ, the mean must sit at least z_* standard deviations above 0, where $\Pr[\mathcal{N}(0,1) > z_*] = \epsilon$. So the number of samples n needed to resolve a given θ is set by requiring this mean to be sufficiently far from 0:

$$\frac{\sqrt{nF\theta^2}}{2} \geq z_*, \qquad |\theta_{\min}| \sim \frac{2z_*}{\sqrt{nF}}. \tag{7.94}$$

This is statistically optimal: the likelihood ratio test is the optimal test for distinguishing P_θ from P_0, so (up to the constant z_*) no procedure can resolve a smaller θ with n samples.

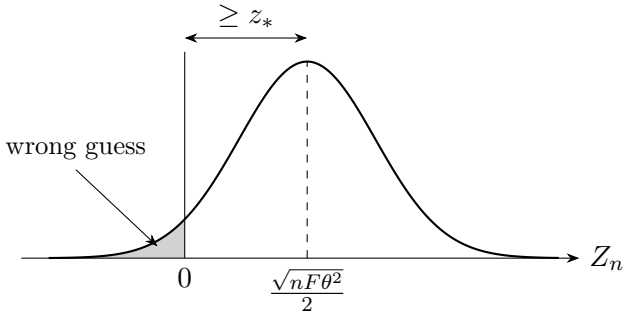


Figure 7.9: Approximate distribution of Z_n when sampling from P_θ: a unit-variance Gaussian centered at $\sqrt{nF\theta^2}/2$. We guess wrong when $Z_n < 0$ (shaded). Keeping this probability small requires the mean to be at least z_* standard deviations from 0.

We can make the same point about *estimating* θ, rather than testing it against θ_0, using the Cramér–Rao bound. Take n i.i.d. samples $\vec{x}$ and an estimator $\hat{\theta}(\vec{x})$. We say $\hat{\theta}$ is a locally unbiased estimator if

$$\mathbb{E}_{\vec{x}\sim P_\theta}\, \hat{\theta}(\vec{x}) = \theta + O((\theta - \theta_0)^2). \tag{7.95}$$

We want to bound the variance of a locally unbiased estimator using the Fisher information.

Theorem 7.2.1 (Cramér–Rao bound)**.** *If $\hat{\theta}$ is locally unbiased, then*

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nF}. \tag{7.96}$$

Proof. In this proof x stands for all n samples $\vec{x}$. Differentiating the definition of local unbiasedness gives $\left[\partial_\theta \sum_x P_\theta(x)\hat{\theta}\right]_{\theta=\theta_0} = 1$, which holds for every locally unbiased estimator. Since $\sum_x \partial_\theta P_\theta(x) = 0$, we may replace $\hat{\theta}$ by $\hat{\theta} - \theta_0$ here. Then, defining an inner product $\langle A, B\rangle = \sum_x P(x)A(x)B(x)$ and using the Cauchy–Schwarz inequality,

$$\begin{aligned} 1 = 1^2 &= \left(\sum_x P_\theta(x) \frac{(\hat{\theta}-\theta_0)\,\partial_\theta P_\theta(x)}{P_\theta(x)}\right)^2 \\ &\leq \left(\sum_x P_\theta(x)(\hat{\theta}-\theta_0)^2\right)\left(\sum_x P_\theta(x)\left(\frac{\partial_\theta P_\theta(x)}{P_\theta(x)}\right)^2\right) = \mathrm{Var}(\hat{\theta})F_n = \mathrm{Var}(\hat{\theta})\,nF, \end{aligned} \tag{7.97}$$

all evaluated at $\theta = \theta_0$. The first factor is the variance because $\mathbb{E}_{\theta_0}\hat{\theta} = \theta_0$, and the Fisher information of n i.i.d. samples is $F_n = nF$ because $\log P_\theta(\vec{x})$ is a sum of n i.i.d. terms. □

7.3 Quantum Fisher information and the quantum Cramér–Rao bound

In the Ramsey protocol, any measurement in the xy plane has the same Fisher information. Indeed, measuring along an axis at angle α in the equatorial plane gives an outcome with probability $P = \frac{1}{2}(1 + \cos(\Delta T - \alpha))$, and for a two-outcome measurement (7.91) gives

$$F = \frac{(\partial_\Delta P)^2}{P} + \frac{(\partial_\Delta P)^2}{1-P} = \frac{(\partial_\Delta P)^2}{P(1-P)} = \frac{\frac{T^2}{4}\sin^2(\Delta T - \alpha)}{\frac{1}{4}\sin^2(\Delta T - \alpha)} = T^2, \tag{7.98}$$

independent of α. This is the same cancellation we saw in Section 7.1, and the Cramér–Rao bound $\mathrm{Var}(\hat{\Delta}) \geq 1/(MT^2)$ matches $\delta\Delta_{\min} \approx 1/(T\sqrt{M})$. In general, though, the choice of measurement matters.

Say we prepare ρ_θ and ρ_0; let us not worry about how to prepare them. We focus on what measurement to perform to extract information about θ. One tempting definition of the quantum Fisher information is

$$\partial_\theta^2 D(\rho_\theta \| \rho_0)\big|_{\theta=0} = \mathcal{F}_Q, \tag{7.99}$$

which turns out to be a valid quantity but does not have the nice features we want, since ρ_θ and $\partial_\theta \rho_\theta$ need not commute, unlike in the classical case. Defining the derivative in different ways gives many variants of the quantum Fisher information; we will look at one of them.

Define $\mathrm{Multi}_\rho[A] = \frac{1}{2}(\rho A + A\rho)$ and $\mathrm{Divid}_\rho[A] = \mathrm{Multi}_\rho^{-1}[A]$ (which is surprisingly well-defined). Then the *symmetric logarithmic derivative* (SLD) $L_{(\rho,\theta)}$ is

$$L_{(\rho,\theta)} = \mathrm{Divid}_\rho[\partial_\theta \rho_\theta], \tag{7.100}$$

which is morally equivalent to $\partial_\theta \log P_\theta(x)$ (if ρ_θ and $\partial_\theta\rho_\theta$ commute, $L = \rho_\theta^{-1}\partial_\theta\rho_\theta = \partial_\theta \log \rho_\theta$). Implicitly, L satisfies

$$\partial_\theta \rho_\theta = \frac{1}{2}\left(\rho L + L\rho\right), \tag{7.101}$$

and a brute-force calculation shows that L is Hermitian: in the eigenbasis of $\rho_\theta = \sum_i \lambda_i |i\rangle\langle i|$, $L_{ij} = \frac{2(\partial_\theta \rho_\theta)_{ij}}{\lambda_i + \lambda_j}$. Then we can define the SLD quantum Fisher information

$$F_Q = \operatorname{tr}(\rho_\theta L^2). \tag{7.102}$$

Theorem 7.3.1 (Braunstein–Caves)**.**

$$F_Q = \sup_{POVM\ \Pi} F_\Pi, \tag{7.103}$$

where the POVM Π induces classical probabilities $P_\theta(k) = \operatorname{tr}(\rho_\theta \Pi_k)$, and F_Π is their classical Fisher information. The supremum is attained by measuring in the eigenbasis of L.

Proof. Let the POVM be $\Pi = \{\Pi_k\}$, with $\sum_k \Pi_k = I$ and $\Pi_k \geq 0$. Note

$$F_\Pi = \sum_k \operatorname{tr}(\rho_\theta \Pi_k) \left(\frac{\partial_\theta \operatorname{tr}(\rho_\theta \Pi_k)}{\operatorname{tr}(\rho_\theta \Pi_k)} \right)^2, \tag{7.104}$$

and

$$\partial_\theta \operatorname{tr}(\rho_\theta \Pi_k) = \frac{1}{2} \left[\operatorname{tr}(\rho_\theta L \Pi_k) + \operatorname{tr}(\rho_\theta \Pi_k L) \right] = \operatorname{Re}(\operatorname{tr}(\rho_\theta L \Pi_k)) \tag{7.105}$$

(the two traces are complex conjugates of each other), so we get

$$\begin{aligned} F_\Pi &\leq \sum_k \frac{|\operatorname{tr}(\rho_\theta L \Pi_k)|^2}{\operatorname{tr}(\rho_\theta \Pi_k)} = \sum_k \left| \frac{\operatorname{tr}\left((\sqrt{\Pi_k} L \sqrt{\rho_\theta})^\dagger (\sqrt{\Pi_k} \sqrt{\rho_\theta}) \right)}{\sqrt{\operatorname{tr}(\rho_\theta \Pi_k)}} \right|^2 \\ &\leq \sum_k \operatorname{tr}\left((\sqrt{\Pi_k} L \sqrt{\rho_\theta})^\dagger (\sqrt{\Pi_k} L \sqrt{\rho_\theta}) \right) \frac{\operatorname{tr}\left((\sqrt{\Pi_k} \sqrt{\rho_\theta})^\dagger (\sqrt{\Pi_k} \sqrt{\rho_\theta}) \right)}{\operatorname{tr}(\rho_\theta \Pi_k)} \\ &= \sum_k \operatorname{tr}(\Pi_k L \rho_\theta L) = \operatorname{tr}(\rho_\theta L^2) = F_Q, \end{aligned} \tag{7.106}$$

where the first step uses $(\operatorname{Re} z)^2 \leq |z|^2$ and the second is the Cauchy–Schwarz inequality for $\langle A, B \rangle = \operatorname{tr}(A^\dagger B)$. We can figure out when the inequalities are saturated. For the first, $\operatorname{tr}(\rho_\theta L \Pi_k)$ must be real, meaning $\operatorname{Im} \operatorname{tr}(\rho_\theta L \Pi_k) = 0$; for the second, we need $\sqrt{\Pi_k} L \sqrt{\rho_\theta} \propto \sqrt{\Pi_k} \sqrt{\rho_\theta}$. Both are satisfied by choosing Π_k to project onto the eigenspaces of L: then $\Pi_k L = \ell_k \Pi_k$ for a real eigenvalue ℓ_k, so $\sqrt{\Pi_k} L \sqrt{\rho_\theta} = \ell_k \sqrt{\Pi_k} \sqrt{\rho_\theta}$, and $\operatorname{tr}(\rho_\theta L \Pi_k) = \ell_k \operatorname{tr}(\rho_\theta \Pi_k)$ is real. □

We end with a few remarks.

Additivity. For n copies $\rho_\theta^{\otimes n}$, the SLD is $L_n = \sum_{i=1}^n L^{(i)}$, where $L^{(i)}$ is L acting on the ith copy (one checks that $\frac{1}{2}(\rho_\theta^{\otimes n} L_n + L_n \rho_\theta^{\otimes n}) = \partial_\theta \rho_\theta^{\otimes n}$). Since $\operatorname{tr}(\rho_\theta L) = \operatorname{tr}(\partial_\theta \rho_\theta) = 0$, the cross terms in $\operatorname{tr}(\rho_\theta^{\otimes n} L_n^2)$ vanish, so $F_Q(\rho_\theta^{\otimes n}) = n F_Q(\rho_\theta)$. The quantum Fisher information is therefore additive, and measuring each copy separately in the eigenbasis of L already achieves it: for a single parameter, joint measurements across copies do not help.

Quantum Cramér–Rao bound. Combining Theorem 7.3.1 with Theorem 7.2.1, any locally unbiased estimator built from any measurement on $\rho_\theta^{\otimes n}$ satisfies

$$\operatorname{Var}(\hat{\theta}) \geq \frac{1}{n F_Q}. \tag{7.107}$$

Similarly, if we estimate θ from the expectation value of an observable O, the resulting error obeys

$$\frac{\operatorname{Var}_{\rho_\theta}(O)}{(\partial_\theta \operatorname{tr}(\rho_\theta O))^2} \geq \frac{1}{F_Q}, \tag{7.108}$$

no matter which observable O we choose.

Multiple parameters. Here we only discussed estimating a single parameter. With more than one parameter, much of this framework breaks down: the optimal measurements for different parameters (the eigenbases of their SLDs) need not commute, so in general the quantum Cramér–Rao bound cannot be saturated for all parameters at once. One way to approach this is geometric: the Fisher information is a second derivative, i.e. a curvature, and with several parameters it becomes a matrix that defines a metric on the space of states.

8.372 Quantum Information Science III **Fall 2026**

Lecture 8: October 6, 2026

Scribe: Aarush Vailaya *Classical messages over quantum channels*

8.1 Classical messages over quantum channels

We will treat sending classical messages over quantum channels as a goal in and of itself, and as a stepping stone to other goals.

Peter Shor once complained (roughly paraphrased) that he wished the physicists would stop talking about quantum effects in channels, since they were giving one ad hoc correction after another, and would just give the general formula. They were thinking of a communication channel with quantum effects. But now, we think of the channel as naturally working with quantum information, and we want to know its classical capacity.

If $\mathcal{N}$ is a channel, then define $C(\mathcal{N})$ as its classical capacity, meaning the largest rate (in bits per channel use) at which classical messages can be sent with vanishing error in the asymptotic limit; we make this precise in Definition 8.3.1. Throughout, Latin letters (x, y, ...) label orthonormal computational basis states and Greek letters (ρ, σ, ψ, ...) denote general states.

In a classical channel, we have $x \mapsto y$ with probability $N(y \mid x)$. In the quantum language, this is

$$\mathcal{N}(\rho) = \sum_x \langle x|\rho|x\rangle \left(\sum_y N(y \mid x) \, |y\rangle\langle y| \right), \tag{8.109}$$

and this is how we should treat classical channels (like a wire): we could input a superposition of voltages, but the channel would collapse it to zeros and ones.

We specialize to CQ (classical-quantum) channels, meaning classical input x to quantum output ρ_x. These are a special case of general quantum channels; as a quantum channel, a CQ channel is

$$\mathcal{N}(\sigma) = \sum_x \langle x|\sigma|x\rangle \, \rho_x. \tag{8.110}$$

The classical channel (8.109) is the special case in which every $\rho_x = \sum_y N(y \mid x) \, |y\rangle\langle y|$ is diagonal in the same basis; in a CQ channel, the ρ_x can be arbitrary.

The noiseless classical channel has

$$\mathcal{N}(\rho) = \sum_x \langle x|\rho|x\rangle \, |x\rangle\langle x| = \sum_x |x\rangle\langle x| \, \rho \, |x\rangle\langle x| \, . \tag{8.111}$$

If the channel acts on a qubit, the noiseless classical channel needs at most 2 Kraus operators ($|0\rangle\langle 0|$ and $|1\rangle\langle 1|$), while a general classical channel needs at most 4 ($\sqrt{N(y \mid x)} \, |y\rangle \langle x|$ for $x, y \in \{0, 1\}$). In a CQ channel, each x could in general need several Kraus operators.

Note that we cannot use a single Kraus operator for each x, as in

$$\mathcal{N}(\sigma) = \sum_x (\sqrt{\rho_x} \otimes \langle x|) \, \sigma \, (\sqrt{\rho_x} \otimes |x\rangle). \tag{8.112}$$

Each Kraus operator E_j maps a pure input to a pure (unnormalized) output, so on a pure input a channel's output is a sum of one rank-1 term per Kraus operator. Its rank (the number of

nonzero eigenvalues) is therefore at most the number of Kraus operators. Since the input $|x\rangle$ must produce ρ_x, a CQ channel needs at least $\operatorname{rank}\rho_x$ Kraus operators for each x; writing $\rho_x = \sum_k \lambda_{x,k} |\psi_{x,k}\rangle\langle\psi_{x,k}|$, the operators $\sqrt{\lambda_{x,k}}\,|\psi_{x,k}\rangle\langle x|$ work. If the input is 1-dimensional but the output is the 100-dimensional maximally mixed state, then we need 100 Kraus operators.

For classical channels, 4 Kraus operators on a qubit is not overkill, because the completely depolarizing channel $\mathcal{D}(\rho) = I/2$ needs at least 4 Kraus operators in

$$\mathcal{D}(\rho) = \sum_{j=1}^{m} E_j \rho E_j^\dagger. \tag{8.113}$$

Some examples are the Paulis scaled by $\frac{1}{2}$, $E_j = \sigma_j/2$, or $E_{ab} = |a\rangle\langle b|/\sqrt{2}$ for $a, b \in \{0,1\}$. To see that $m \geq 4$, use the rank bound above. Note that $\mathrm{id} \otimes \mathcal{D}$ has the same number of Kraus operators, $I \otimes E_j$. Putting a Bell pair $|\Phi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$ into $\mathrm{id} \otimes \mathcal{D}$ gives $\operatorname{tr}_B |\Phi\rangle\langle\Phi| \otimes \frac{I}{2} = \frac{I}{2} \otimes \frac{I}{2} = I/4$, which has rank 4, so $m \geq 4$.

8.2 Superdense coding

We take a little detour to superdense coding. With $|\Phi\rangle = \frac{|00\rangle+|11\rangle}{\sqrt{2}}$ as above, define $|\Psi_j\rangle = (\sigma_j \otimes I)|\Phi\rangle$ for the Paulis $\sigma_0 = I, \sigma_1, \sigma_2, \sigma_3$. These are orthonormal, since

$$\langle\Psi_i|\Psi_j\rangle = \frac{1}{2}\operatorname{tr}\sigma_i\sigma_j = \delta_{ij}. \tag{8.114}$$

This leads to a protocol called superdense coding, where

$$1 \text{ qubit} + 1 \text{ ebit} \geq 2 \text{ cbits}. \tag{8.115}$$

Here a cbit is the ability to transmit 1 classical bit, an ebit is one shared Bell pair, and a qubit is the ability to transmit 1 qubit. Later today, we will see 1 qubit $\geq$ 1 cbit: the ability to send a quantum bit is at least as strong as the ability to send a classical bit (trivially, send $|0\rangle$ or $|1\rangle$). But the qubit has continuous degrees of freedom, so maybe we can send more than one bit... It turns out that Holevo's theorem states we cannot send more than 1 bit.

So on its own, a qubit cannot carry more than one classical bit, but if Alice and Bob share a bit of entanglement, they can send information at twice the rate. Alice and Bob share $|\Phi\rangle$, and Alice applies the Pauli σ_j to her side of the state, creating one of the four orthogonal states $|\Psi_j\rangle$. She sends her part of the state to Bob (so the entanglement is gone, as it is all in Bob's lab), and now he has 2 qubits in one of 4 orthogonal states, so he can recover j exactly. Note we only send one qubit through a perfect qubit channel, but we use up the entanglement between them.

This turns out to be tight; we cannot send one qubit and expend more entanglement to send more than 2 classical bits. Teleportation is similar:

$$2 \text{ cbits} + 1 \text{ ebit} \geq 1 \text{ qubit}. \tag{8.116}$$

Suppose Alice and Bob share unlimited entanglement. Then teleportation and superdense coding together give 1 qubit = 2 cbits. If either inequality were not tight (say 1 qubit plus ebits gave 3 cbits), then alternating the two protocols would turn 2 cbits into 3, then 3 into more, and so on, letting them send unboundedly many cbits from 2, which is impossible. Thus, given unlimited ebits, 2 cbits equal 1 qubit. You also cannot improve on the amount of entanglement used, but that is a bit subtle.

8.3 The HSW theorem

Let us recall what the capacity of a channel means.

Definition 8.3.1. *Let $\mathcal{N}$ be a CQ channel with input alphabet $[d_A]$ and outputs ρ_x that are $d_B \times d_B$ density matrices. We say (M, n, ϵ) is achievable if there exist an encoder $E : [M] \to [d_A]^n$ and operators $D_1, \dots, D_M$ with $D_m \geq 0$ and $\sum_m D_m \leq I$ such that for all $m \in [M]$,*

$$\operatorname{tr}\left[D_m \mathcal{N}^{\otimes n}(|E(m)\rangle\langle E(m)|)\right] \geq 1 - \epsilon. \tag{8.117}$$

So we can send M messages using n uses of the channel, with error probability at most ϵ for every message. The D_m are $d_B^n \times d_B^n$ measurement matrices, acting on Bob's n output systems, where $I - \sum_m D_m$ is a "give up" outcome. The capacity is

$$C(\mathcal{N}) = \lim_{\epsilon \to 0} \lim_{n \to \infty} \max_{M:(M,n,\epsilon) \text{ achievable}} \frac{1}{n} \log M. \tag{8.118}$$

If $M = 2^{nR}$, the above gives the largest achievable rate R.

Holevo is still alive, and the HSW (Holevo–Schumacher–Westmoreland) theorem says the following.

Theorem 8.3.1 (HSW)**.** *For a CQ channel $\mathcal{N} : x \mapsto \rho_x$,*

$$C(\mathcal{N}) = \max_p I(X; Q)_\omega, \tag{8.119}$$

where

$$\omega^{XQ} = \sum_x p(x) \, |x\rangle\langle x|^X \otimes \rho_x^Q. \tag{8.120}$$

The state ω^{XQ} is classical on the X register and quantum on the Q register; $p(x)$ is Alice's input distribution, and ω is the joint state of Alice's input and Bob's output. Like Shannon's theorem, this gives us the classical capacity of the channel in terms of a mutual information.

There are a few ways of understanding this. One way is

$$I(X; Q) = S(Q) - S(Q \mid X) = S\Big(\sum_x p_x \rho_x\Big) - \sum_x p_x S(\rho_x), \tag{8.121}$$

where $S(Q \mid X) = \sum_x p_x S(\rho_x)$ because X is classical: $S(XQ) = H(p) + \sum_x p_x S(\rho_x)$. Note this is nonnegative due to concavity. Let $\rho = \sum_x p_x \rho_x$ denote the average state. In older literature, the quantity $S(\rho) - \sum_x p_x S(\rho_x)$ is called the Holevo quantity χ, but people later realized this is mutual information.

For example, consider the three "Mercedes" states (Figure 8.10). We want to transmit through the CQ channel $x \mapsto \rho_x$ that outputs these states. Since these states are not orthogonal, from one use of the channel, Bob cannot get a perfect bit. It looks like the capacity should be less than one, but HSW says the capacity is one. Suppose we put in $p = (1/3, 1/3, 1/3)$; the average density matrix is $\rho = I/2$. Then, $S(\rho) = 1$ as ρ is maximally mixed. Since each ρ_x is pure and has entropy 0, $\sum_x p_x S(\rho_x) = 0$, so $I(X; Q) = 1$ (and no input distribution does better, since $S(\rho) \leq 1$ for a qubit). So asymptotically this channel is as good as a noiseless one-bit channel.

There is also a relation to relative entropy. Recall $\rho = \sum_x p_x \rho_x$, and look at $D(\rho_x \| \rho) = -S(\rho_x) - \operatorname{tr} \rho_x \log \rho$. Then

$$\sum_x p_x D(\rho_x \| \rho) = -\sum_x p(x) S(\rho_x) + S(\rho) = I(X; Q). \tag{8.122}$$

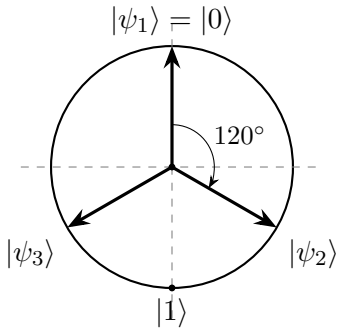


Figure 8.10: The three "Mercedes" states, drawn in a cross-section of the Bloch sphere through the poles: pure qubit states whose Bloch vectors are 120° apart, with $|0\rangle$ at 12:00. They are not orthogonal, but their uniform average is $I/2$.

In the classical case, relative entropy was a measure of surprise. Here, it measures how surprising it is to see ρ_x if we were expecting ρ, and averaging it gives us the mutual information. This is exactly how channels work: Bob expects the average output ρ (as he knows the distribution of symbols), but he receives the output ρ_x for the particular symbol sent, which gives some surprise, and on average this surprise is the mutual information.

This will relate to the seemingly unrealistic error conditions in hypothesis testing (Lecture 6), where we wanted constant error of one type and exponentially decreasing error of the other type.

8.4 Proof of HSW achievability

We now show that we can send at any rate $R < I(X;Q)_\omega$. Fix some distribution p (which might be the argmax for $C(\mathcal{N})$, but it can be any distribution). Sample the codewords

$$x^n(1), \dots, x^n(M) = E(1), \dots, E(M) \sim p^n, \tag{8.123}$$

which are i.i.d. strings: for each codeword we sample n i.i.d. symbols from p. Then

$$\sigma_m = \mathcal{N}^{\otimes n}(|x^n(m)\rangle\langle x^n(m)|) = \rho_{x_1(m)} \otimes \cdots \otimes \rho_{x_n(m)}. \tag{8.124}$$

So every message m corresponds to a different state σ_m (in the Mercedes example, one message might encode to 12:00 $\otimes$ 4:00 $\otimes$ 8:00 $\otimes$ 4:00). If $t = \text{type}(x^n(m))$, then σ_m is a permuted version of $\rho_1^{\otimes nt_1} \otimes \rho_2^{\otimes nt_2} \otimes \cdots$. Averaging over the codeword, each position is independently ρ_x with probability $p(x)$, so

$$\mathbb{E}_{x^n(m)} \sigma_m = \rho^{\otimes n}. \tag{8.125}$$

The following is not a necessary part of the proof, but motivates it. Consider $D(\sigma_m \| \rho^{\otimes n})$. To evaluate the relative entropy we can do it copy by copy, to get

$$D(\sigma_m \| \rho^{\otimes n}) = \sum_x n t_x D(\rho_x \| \rho) \approx n I(X;Q)_\omega, \tag{8.126}$$

by (8.122), if $t \approx p$ (which, as we showed in a problem set a while back, is very likely).

We define Π_m as the conditionally typical projector for σ_m. For some permutation π of the n systems, it equals

$$\Pi_m = \pi\Big[\bigotimes_x \Pi_{\rho_x,\delta}^{nt_x}\Big]\pi^\dagger. \tag{8.127}$$

This projector does what we want for hypothesis testing, because

$$\operatorname{tr}(\sigma_m \Pi_m) \geq 1-\epsilon, \tag{8.128}$$

by inheriting the guarantee for each typical projector. In particular, $\operatorname{tr}(\sigma_m\Pi_m) = \prod_x \operatorname{tr}\big(\rho_x^{\otimes nt_x}\Pi_{\rho_x,\delta}^{nt_x}\big)$. Each factor is at least $1-\epsilon'$ once nt_x is large (true for every x with $p(x)>0$, since $t \approx p$), and there are at most d_A factors, a number that does not grow with n. So the product is at least $1-d_A\epsilon'$, which we call $1-\epsilon$. And[3]

$$\operatorname{tr}(\rho^{\otimes n}\Pi_m) \leq \exp\Big(-n\sum_x t_x D(\rho_x\|\rho)\Big) \approx \exp(-nI(X;Q)). \tag{8.129}$$

This is the error which we want to have an exponential rate of decay, and the decay rate gives us our communication rate. Why?

Classically, Bob would use sequential decoding: Alice sends σ_m, and Bob measures $\{\Pi_1, I-\Pi_1\}$, then $\{\Pi_2, I-\Pi_2\}$, and so on. He basically asks whether the message is 1; if so he stops and returns 1, otherwise he asks whether it is 2, etc. This works because if the message is 17, then on the 17th measurement Bob wrongly gets $I-\Pi_{17}$ only with constant probability ϵ. For the other measurements, Bob wrongly accepts with at most exponentially small probability: the codeword $x^n(m')$ is independent of $x^n(m)$, so by (8.125), averaged over the codebook σ_m looks like $\rho^{\otimes n}$ to $\Pi_{m'}$, and (8.129) applies. By the classical union bound, this finishes the classical proof:

$$\begin{aligned}
\mathbb{E}_{\text{codebook}} \Pr\big[\text{Bob outputs } m' \neq m \mid \text{Alice sends } m\big] &\leq \mathbb{E}_{\text{codebook}} \operatorname{tr}\big((I-\Pi_m)\sigma_m\big) \\
&\quad + \sum_{m'\neq m} \mathbb{E}_{\text{codebook}} \operatorname{tr}(\Pi_{m'}\sigma_m) \\
&\leq \epsilon + \sum_{m'\neq m} \mathbb{E}_{\text{codebook}} \operatorname{tr}(\Pi_{m'}\rho^{\otimes n}) \\
&\lesssim \epsilon + (M-1)\exp(-nI(X;Q)) \\
&\approx \epsilon + \exp(n(R-I(X;Q))),
\end{aligned} \tag{8.130}$$

and the second term goes to 0 when $R < I(X;Q)$.

Why expect this not to work quantumly? Well, gentle measurement is strong, but we would not expect it to be strong enough, since Bob makes exponentially many measurements, each slightly disturbing the state. Essentially, we have to combine $\Pi_1, \ldots, \Pi_M$ into some measurement. The old answer was the PGM (pretty good measurement), where

$$D_m = A^{-1/2}\Pi_m A^{-1/2}, \qquad A = \sum_m \Pi_m; \tag{8.131}$$

analyzing it is a bit annoying, but it works. There is a simpler argument, where we just show sequential measurement works.

To do this we use the quantum version of the union bound, called the non-commutative union bound.

[3] This bound needs care when ρ_x does not commute with ρ. With the plain typical projectors it can fail: for $\rho_0 = |0\rangle\langle 0|$, $\rho_1 = |+\rangle\langle +|$ and $p = (\frac{1}{2}, \frac{1}{2})$, so $\rho = \frac{1}{2}(|0\rangle\langle 0| + |+\rangle\langle +|)$, we have $\Pi_m = \sigma_m$ and $\operatorname{tr}(\rho^{\otimes n}\Pi_m) = (3/4)^n \approx 2^{-0.42n}$, while $I(X;Q) = S(\rho) \approx 0.60$. The fix is to build Π_m from the hypothesis-testing operators $\Pi_{\rho_x\|\rho}\Pi_{\rho_x}\Pi_{\rho_x\|\rho}$ from the proof of quantum Stein's lemma (Lecture 6), or to sandwich with the typical projector of $\rho^{\otimes n}$.

Lemma 8.4.1 (Non-commutative union bound)**.** *Let $\omega \geq 0$ with $\operatorname{tr}\omega \leq 1$, and let $P_1, \ldots, P_L$ be projectors. Then*

$$\operatorname{tr}[\omega] - \operatorname{tr}[P_L \cdots P_1 \omega P_1 \cdots P_L] \leq 2\sqrt{\sum_{i=1}^{L} \operatorname{tr}[(I - P_i)\omega]}. \tag{8.132}$$

We should think of ω as a density matrix (the lemma is just written in more generality), and of P_i as the outcome we hope for in measurement i (for Bob, the "no" outcome $I - \Pi_{m'}$ for the wrong messages m'). So the chance that at least one measurement gives a bad outcome is at most twice the square root of the sum of the individual probabilities of the bad outcomes.

Now apply Lemma 8.4.1 with $\omega = \sigma_m$, $P_{m'} = I - \Pi_{m'}$ for $m' < m$, and $P_m = \Pi_m$. Bob's error probability is at most $2\sqrt{\operatorname{tr}((I - \Pi_m)\sigma_m) + \sum_{m'<m} \operatorname{tr}(\Pi_{m'}\sigma_m)}$, and averaging over the codebook (using that the square root is concave) gives

$$\mathbb{E}_{\text{codebook}} \Pr\big[\text{Bob outputs } m' \neq m \mid \text{Alice sends } m\big] \lesssim 2\sqrt{\epsilon + \exp(n(R - I(X;Q)))}. \tag{8.133}$$

This is the classical union bound (8.130), up to a factor of 2 and a square root, which do not change the achievable rate.

8.372 Quantum Information Science III **Fall 2026**

Lecture 9: October 8, 2026

Scribe: Aarush Vailaya *HSW converse*

9.1 Converse for the HSW theorem

Last time we proved achievability for the HSW theorem, which says that the capacity of a CQ channel $\mathcal{N}$ is

$$C(\mathcal{N}) = \max_p I(X;Q)_\omega, \qquad \omega = \sum_x p(x) \, |x\rangle\langle x|^X \otimes \mathcal{N}(|x\rangle\langle x|)^Q, \tag{9.134}$$

where ω is a CQ state. Today, we will prove the converse.

Step 1: setup. The high level idea is to identify a Markov chain. In particular, Alice has a message M which she encodes in X^n, which goes through the channel to produce Q^n, which Bob decodes to get $\hat{M}$:

$$M - X^n - Q^n - \hat{M}. \tag{9.135}$$

Note $M, \hat{M} \in \{0,1\}^{nR}$, and our error condition is $\Pr\left[M \neq \hat{M}\right] \leq \epsilon$ for all M (where the probability is over the protocol and the channel). Thinking of this as a Markov structure, the only way M can affect the quantum state Q^n is via X^n.

Step 2: put everything in terms of entropies. Note $H(M) \leq nR$, and we will choose M to be uniformly distributed so $H(M) = nR$. We can do this because we are proving the converse: a code with error at most ϵ for every message also has error at most ϵ when the message is uniformly random, so showing that no code does well on average is a stronger statement than showing that no code does well in the worst case. We also want to put the error probability in terms of entropy, which we do via Fano's inequality. The idea is that if we know $\hat{M}$, we pretty much know M, since $\hat{M} = M$ with probability at least $1 - \epsilon$. For entropies, we state this using conditional entropy.

Lemma 9.1.1 (Fano's inequality)**.**

$$H(M \mid \hat{M}) \leq 1 + \epsilon nR. \tag{9.136}$$

Proof. Let $d = 2^{nR}$ be the number of possible messages. Fix a value of $\hat{M}$ and let $p = \Pr\left[\cdot \mid \hat{M}\right]$, the distribution of M conditioned on this value. Then,

$$p = (1-\delta)1_{\hat{M}} + \delta q, \tag{9.137}$$

where $\delta = \Pr\left[M \neq \hat{M} \mid \hat{M}\right]$ is the error probability given $\hat{M}$ (which on average over $\hat{M}$ is at most ϵ), and q is a distribution with $q(\hat{M}) = 0$ that is possibly supported everywhere else. We get

$$\begin{aligned} H(p) &= -(1-\delta)\log(1-\delta) - \sum_{x \neq \hat{M}} \delta q(x) \log \delta q(x) \\ &= H_2(\delta) + \delta H(q) \\ &\leq 1 + \delta \log(d-1), \end{aligned} \tag{9.138}$$

where $H_2(p) = -p\log p - (1-p)\log(1-p) \leq 1$ is the binary entropy. The idea is that $H_2(\delta)$ is the bit of information on whether we are at $\hat{M}$, and $\delta H(q)$ is the information on which wrong message we have, which is needed only with probability δ. Averaging over $\hat{M}$,

$$H(M \mid \hat{M}) = \mathbb{E}_{\hat{M}} H(p) \leq 1 + \epsilon \log(d-1) \leq 1 + \epsilon n R. \tag{9.139}$$

□

Fano's inequality is what we need here, but in general, one often wants another inequality:

Theorem 9.1.1 (Fannes' inequality)**.**

$$|H(p) - H(q)| \leq H_2(\epsilon) + \epsilon \log d, \tag{9.140}$$

where $\epsilon = \frac{1}{2}\|p-q\|_1$*. The quantum version is*

$$|S(\rho) - S(\sigma)| \leq H_2(\epsilon) + \epsilon \log d, \tag{9.141}$$

where $\epsilon = \frac{1}{2}\|\rho - \sigma\|_1$*.*

Rough intuition: moving a probability mass of ϵ (in trace distance) can change the entropy by about $\epsilon \log d$. The $H_2(\epsilon)$ term also vanishes as $\epsilon \to 0$, but only like $\epsilon \log(1/\epsilon)$ rather than like ϵ. Since $H_2(\epsilon) \leq 1$ while $\log d$ is typically proportional to n, the $H_2(\epsilon)$ term is usually not the relevant one.

9.2 Conditional mutual information

Step 3: the Markov condition. There is one more thing we can put in terms of entropies, which is the Markov condition: M determines X, X determines Q, and Q determines $\hat{M}$. This is a much deeper step. Recall the conditional mutual information (CMI),

$$\begin{aligned} I(X;Y \mid Z) &= \sum_z p_z(z)\, I(X;Y \mid Z=z) && \text{(classical only)} \\ &= S(X \mid Z) + S(Y \mid Z) - S(XY \mid Z) \\ &= S(XZ) + S(YZ) - S(Z) - S(XYZ) \\ &= I(X;YZ) - I(X;Z). \end{aligned} \tag{9.142}$$

The first expression only makes sense classically; the others work for quantum states too. Cryptographically, this is intuitively the information that X and Y share that Z does not know about (Figure 9.11). The CMI is nonnegative: classically it is an average of mutual informations, and quantumly $I(A;B \mid C) \geq 0$ is strong subadditivity, which follows from monotonicity of relative entropy (Lecture 6).

We have the chain rule for mutual information,

$$I(X;YZ) = I(X;Z) + I(X;Y \mid Z), \tag{9.143}$$

meaning

$$I(X;Y_1 \cdots Y_n) = I(X;Y_1) + I(X;Y_2 \mid Y_1) + \cdots + I(X;Y_n \mid Y_1 \cdots Y_{n-1}). \tag{9.144}$$

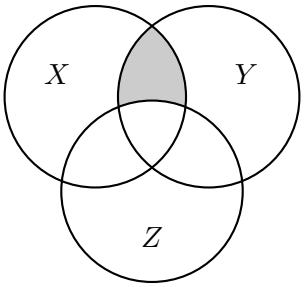


Figure 9.11: Venn diagram for three random variables. The shaded region is $I(X;Y \mid Z)$, the information shared by X and Y that Z does not know about.

One way this is useful is when we think about when entropy is 0. If a quantum state has 0 entropy, it is pure. If a bipartite state has 0 mutual information, then $\rho_{AB} = \rho_A \otimes \rho_B$. Altogether,

$$\begin{aligned} S(A) = 0 &\Leftrightarrow \rho_A \text{ is pure,} \\ I(A;B) = 0 &\Leftrightarrow \rho_{AB} = \rho_A \otimes \rho_B, \\ H(X \mid Y) = 0 &\Leftrightarrow \text{classically, } X \text{ is a deterministic function of } Y, \\ I(X;Y \mid Z) = 0 &\Leftrightarrow X - Z - Y \text{ form a Markov chain.} \end{aligned} \tag{9.145}$$

One way to denote the Markov chain claim is

$$p(x,y,z) = p(z)\, p(x \mid z)\, p(y \mid z). \tag{9.146}$$

This is the entropic condition describing the Markov chain process from above. It is also robust, in that the conditional mutual information is equal to the relative entropy to the nearest Markov chain (classically):

$$I(X;Y \mid Z)_p = \min_{q:\ X-Z-Y \text{ Markov}} D(p \,\|\, q), \tag{9.147}$$

with the minimum attained at $q(x,y,z) = p(z)\, p(x \mid z)\, p(y \mid z)$. In the quantum case, $I(A;B \mid C) = 0$ implies ρ is a quantum Markov state (which we will learn about in 8 or so lectures).

Note conditioning classically will reduce entropy, but can increase or decrease mutual information. Suppose $Y = X + Z$, where X and Z are independent uniform bits and the sum is mod 2; then $I(X;Y) = 0$ but $I(X;Y \mid Z) = 1$, so conditioning on Z increases the mutual information.

One more operational way to phrase conditional mutual information is the data-processing inequality.

Lemma 9.2.1 (Data-processing inequality)**.** *If $X - Y - Z$ is a Markov chain, i.e. $I(X;Z \mid Y) = 0$, then*

$$I(X;Z) \leq I(X;Y). \tag{9.148}$$

Intuitively, in a game of telephone, each person knows at most as much about the original message as the neighbor they heard it from.

Proof. We have $I(X;Z) = I(X;YZ) - I(X;Y \mid Z)$ and $I(X;Y) = I(X;YZ) - I(X;Z \mid Y)$. Then,

$$I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y) = I(X;Y \mid Z), \tag{9.149}$$

which is always nonnegative. □

All of these identities, and the data-processing inequality, work quantumly too, except the first (classical-only) expression for the CMI in (9.142).

9.3 Warmup: classical converse

Step 4: the classical converse. Classically, $M - X^n - Y^n - \hat{M}$ (Figure 9.12), and note

$$I(M;\hat{M}) = H(M) - H(M \mid \hat{M}) \geq (1-\epsilon)nR - 1 \tag{9.150}$$

by Fano's inequality. Then,

$$\begin{aligned}(1-\epsilon)nR - 1 &\leq I(M;\hat{M}) \\ &\leq I(X^n;Y^n) && \text{data processing} \\ &\leq \sum_{j=1}^{n} I(X_j;Y_j) && (*)\text{, not justified yet} \\ &\leq nC && \text{definition of } C,\end{aligned} \tag{9.151}$$

from which we get $R \leq \frac{C}{1-\epsilon}$, and as $\epsilon \to 0$, we get $R \leq C$. This is the converse of Shannon's theorem. Note we are not assuming the Xs are independent.

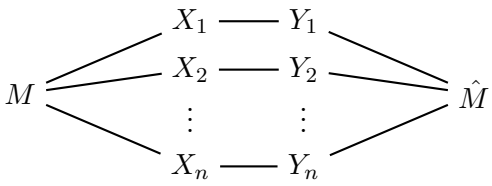


Figure 9.12: The Markov structure of a classical code: the message M determines all the inputs X_j, each output Y_j depends only on X_j, and Bob's guess $\hat{M}$ is computed from all the Y_j.

Let us now do the additivity step, marked $(*)$ above. Note $I(X^n;Y^n) = H(Y^n) - H(Y^n \mid X^n)$ and[4]

$$H(Y^n \mid X^n) = \sum_{j=1}^{n} H(Y_j \mid X^n Y_1 \cdots Y_{j-1}) = \sum_{j=1}^{n} H(Y_j \mid X_j). \tag{9.152}$$

For the second equality, look at the Markov structure in Figure 9.12. Here, Y_j only depends on X_j, so even if the other variables are correlated with Y_j, these correlations are all mediated by X_j. To use this, we phrase it in terms of conditional mutual information, so that we can use the Markov claim. For any A, B, C,

$$H(A \mid B) - H(A \mid BC) = H(AB) - H(B) - H(ABC) + H(BC) = I(A;C \mid B). \tag{9.153}$$

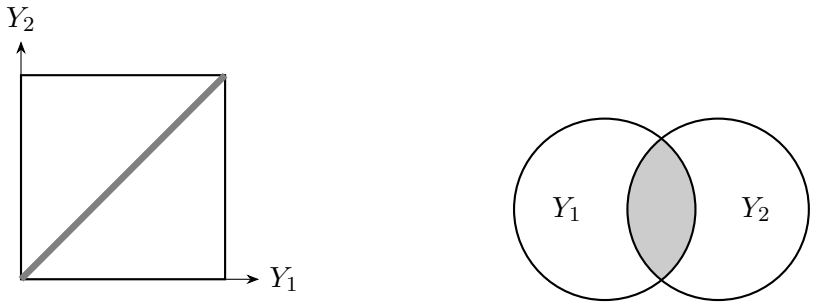
With $A = Y_j$, $B = X_j$ and $C = X_{\neq j} Y_1 \cdots Y_{j-1}$, this is zero by the Markov chain $Y_j - X_j - X_{\neq j} Y_1 \cdots Y_{j-1}$ (the channel acts on each use separately).

For $H(Y^n)$, note $H(Y^n) \leq \sum_{j=1}^{n} H(Y_j)$ by subadditivity (the entropy of the whole is at most the sum of the entropies of the parts). For example, take $Y_1 = Y_2$ uniform on d values, so the joint distribution is spread along the diagonal of the square in Figure 9.13. Each Y_j alone is uniform, with entropy $\log d$, as much as if the whole square were filled, but the correlation means $H(Y_1 Y_2) = \log d < 2 \log d$. We can also see this from the Venn diagram picture: the area of the

[4]The first equality is the chain rule for entropy. To check it, look at $n = 3$: $H(Y_1 Y_2 Y_3 \mid X^n) = H(Y_1 \mid X^n) + H(Y_2 \mid X^n Y_1) + H(Y_3 \mid X^n Y_1 Y_2)$, which follows by writing each term as $H(AB) - H(B)$ and telescoping.

union is at most the area of one plus the area of the other, because of the overlap. Putting these two together gives us the desired statement:

$$I(X^n;Y^n) = H(Y^n) - \sum_{j=1}^{n} H(Y_j \mid X_j) \le \sum_{j=1}^{n} \big(H(Y_j) - H(Y_j \mid X_j)\big) = \sum_{j=1}^{n} I(X_j;Y_j). \tag{9.154}$$



(a) $Y_1 = Y_2$ (b) Venn diagram

Figure 9.13: Subadditivity, $H(Y_1Y_2) \le H(Y_1) + H(Y_2)$. (a) If $Y_1 = Y_2$ is uniform, the joint distribution lies on the diagonal: each marginal is uniform, but the joint entropy is only $\log d$. (b) The joint entropy is the area of the union, which is less than the sum of the two areas by the overlap $I(Y_1;Y_2)$ (shaded).

9.3.1 HSW converse

This also works for CQ channels, by replacing all the Ys with Qs. We get

$$\begin{aligned} I(X^n;Q^n) &= S(Q^n) - S(Q^n \mid X^n) \\ &\le \sum_j \big(S(Q_j) - S(Q_j \mid X_j)\big) \\ &\le nC. \end{aligned} \tag{9.155}$$

The CQ structure is important, since conditioning on a classical variable has a lot of nice properties, letting us do the above (in particular, given $X^n = x^n$ the output is the product state $\rho_{x_1} \otimes \cdots \otimes \rho_{x_n}$, so $S(Q^n \mid X^n) = \sum_j S(Q_j \mid X_j)$). Putting this into the classical argument gives the converse:

$$(1-\epsilon)nR - 1 \le I(M;\hat{M}) \le I(X^n;Q^n) \le \sum_{j=1}^{n} I(X_j;Q_j) \le nC(\mathcal{N}). \tag{9.156}$$

As in (9.151), the second inequality is data processing, now quantum for Bob's measurement of Q^n.

9.4 General channels

Now we want to discuss general quantum channels. It turns out working through CQ channels is a large portion of the work.

Definition 9.4.1. *Define*

$$\chi(\mathcal{N}) = \sup_{p_x,\varphi_x} I(X;Q)_\omega, \qquad \omega = \sum_x p_x \left|x\right\rangle\!\left\langle x\right|^X \otimes \mathcal{N}(\varphi_x)^Q. \tag{9.157}$$

Compared to a CQ channel, Alice now gets to choose the input states φ_x as well as their probabilities p_x. In a CQ channel we have a canonical basis, but in a general channel that is part of the choice. In principle, Alice can input mixed states, but there is no benefit to that (she might as well decompose the mixed state into pure states and keep that as part of her label x). Thus, we can assume φ_x is pure.

It's worth taking a step back to think of the value of Shannon's theory. Shannon took a blockwise asymptotic question and turned it into something single-letter: a maximization over the input to a single use of the channel. If he had turned one n-copy question into another n-copy question, it would have been less interesting. Since the mutual information is concave in the input distribution, this maximization is a convex optimization problem, solvable in time polynomial in the alphabet size. The same holds for the HSW formula for CQ channels. For general channels, $\chi(\mathcal{N})$ also optimizes over the input states φ_x, and this is the hard part: approximating $\chi(\mathcal{N})$ to within $1/\text{poly}(d)$ is NP-complete.[5]

The actual capacity is

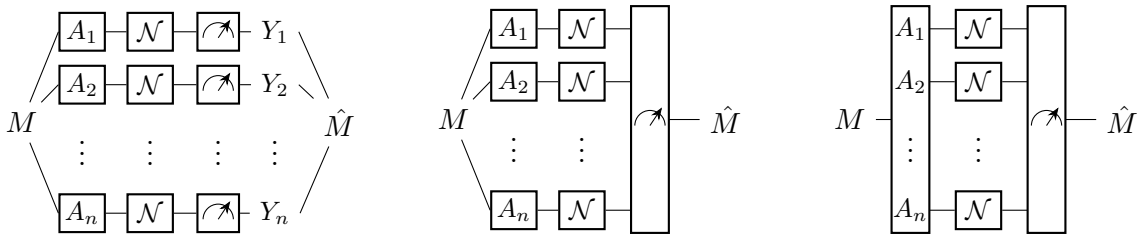
$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n}\chi(\mathcal{N}^{\otimes n}). \tag{9.158}$$

9.4.1 Three settings

There are three settings to consider (Figure 9.14). In the first, Alice's message M determines some product state input on $A_1, \ldots, A_n$, Bob measures each channel output separately to get $Y_1, \ldots, Y_n$, and these all relate to $\hat{M}$. Think of this as a photonic communication link. It is hard to entangle photons, so Bob is measuring them one at a time.

To be a little more generous, in the second setting Alice is still sending unentangled states, but Bob makes a single collective measurement on all the outputs and outputs $\hat{M}$.

Finally, assuming we care about theoretical limits, Alice prepares some possibly entangled state from M, sends it through the n uses of the channel, then Bob measures the whole thing, and gets out $\hat{M}$.



(a) $I_{acc}(\mathcal{N})$ (b) $\chi(\mathcal{N})$ (c) $C(\mathcal{N})$

Figure 9.14: Three settings for sending a classical message M through n uses of a quantum channel $\mathcal{N}$. (a) Product inputs, with each output measured separately; the best rate is the accessible information $I_{acc}(\mathcal{N})$. (b) Product inputs with one collective measurement; the best rate is $\chi(\mathcal{N})$. (c) Inputs possibly entangled across channel uses, with a collective measurement; the best rate is the capacity $C(\mathcal{N})$.

It turns out $\chi(\mathcal{N})$ is the capacity for the middle one, since χ optimizes over the input to a single channel use. Our protocol for achieving this involved sequential decoding, which required

[5] To relate an optimization problem to NP, we turn it into a yes/no question (is the optimum above some threshold?), allowing some error, so hardness depends on the precision. For example, for Max-Cut, the Goemans–Williamson algorithm achieves approximation ratio about 0.878 in polynomial time, and doing better is hard assuming the Unique Games Conjecture.

a block-wise measurement; measuring each output individually only achieves I_{acc}, which can be smaller. The third case is the full capacity $C(\mathcal{N})$, given by (9.158), since we can group together multiple uses of the channel and entangle the inputs across them. Applying the result for the second setting to blocks of k channel uses, i.e. to $\mathcal{N}^{\otimes k}$, achieves rate $\frac{1}{k}\chi(\mathcal{N}^{\otimes k})$, which approaches $C(\mathcal{N})$ as $k \to \infty$; making this rigorous requires some care with the order of limits.

The best rate in the first setting is called the accessible information $I_{acc}(\mathcal{N})$. In general $I_{acc}(\mathcal{N}) \leq \chi(\mathcal{N}) \leq C(\mathcal{N})$, and both inequalities can be strict. For instance, for the Mercedes states from the previous lecture, one can non-trivially show that I_{acc} is less than 1 while $\chi(\mathcal{N}) = 1$. $C(\mathcal{N})$ can differ from $\chi(\mathcal{N})$ because χ is superadditive:

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2), \tag{9.159}$$

since product inputs are allowed, and the inequality can be strict because Alice can entangle the inputs to the two channels. This was called the "additivity problem" before.

8.372 Quantum Information Science III **Fall 2026**

Lecture 10: October 15, 2026

Scribe: Aarush Vailaya *Applications of the HSW converse*

10.1 Additivity of χ

We previously discussed

$$\chi(\mathcal{N}) = \max_{\{p_i,\rho_i\}} S\Big(\sum_i p_i \mathcal{N}(\rho_i)\Big) - \sum_i p_i S(\mathcal{N}(\rho_i)) = \max I(X;Q)_\omega, \tag{10.160}$$

where

$$\omega = \sum_i p_i \left|i\right\rangle\!\left\langle i\right|^X \otimes \mathcal{N}(\rho_i)^Q. \tag{10.161}$$

We also discussed

$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n}\chi(\mathcal{N}^{\otimes n}) \geq \chi(\mathcal{N}), \tag{10.162}$$

where the inequality is true since

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2), \tag{10.163}$$

since product inputs to $\mathcal{N}_1 \otimes \mathcal{N}_2$ achieve the right-hand side, while on the left we can also use entanglement. It turns out there are cases where this inequality is strict, but there are many cases where $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) = \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2)$:

1. $\mathcal{N}_1$ is entanglement-breaking or QCQ, meaning it measures the input with a POVM $\{M_k\}$ and then prepares a state σ_k depending on the outcome k (quantum $\to$ classical $\to$ quantum):

$$\mathcal{N}_1(\rho) = \sum_k \mathrm{tr}[\rho M_k]\sigma_k. \tag{10.164}$$

2. $\mathcal{N}_1$ is depolarizing, meaning

$$\mathcal{N}_1(\rho) = (1-p)\rho + p\frac{I}{d}. \tag{10.165}$$

3. $\mathcal{N}_1$ is an erasure channel, meaning

$$\mathcal{N}_1(\rho) = (1-p)\rho + p\left|e\right\rangle\!\left\langle e\right|. \tag{10.166}$$

4. $\mathcal{N}_1$ is a unital qubit channel, meaning $\mathcal{N}_1(I) = I$.

   Channels are always trace preserving, and $\mathcal{N}$ is unital if and only if $\mathcal{N}^\dagger$ is trace preserving, where $\mathcal{N}^\dagger$ is defined by conjugating all the Kraus operators: if $\mathcal{N}(\rho) = \sum_k A_k \rho A_k^\dagger$, then $\mathcal{N}^\dagger(Y) = \sum_k A_k^\dagger Y A_k$. Equivalently, $\mathcal{N}^\dagger$ is the channel acting on observables (the Heisenberg picture), defined by $\mathrm{tr}[Y\mathcal{N}(\rho)] = \mathrm{tr}[\mathcal{N}^\dagger(Y)\rho]$ for all ρ and Y. Both conditions say $\sum_k A_k A_k^\dagger = I$.

5. $\mathcal{N}_1$ is a pure-loss bosonic channel. Bosonic means the input and output live in the harmonic oscillator state space, which is infinite-dimensional. We feed the input into one port of a beam splitter and the vacuum into the other, and Bob gets one of the output ports. The beam splitter is
$$U = \exp\big(i\eta(a_1^\dagger a_2 + a_1 a_2^\dagger)\big), \qquad \eta \in \mathbb{R}, \tag{10.167}$$
and tracing out the second output gives
$$\mathcal{N}_1(\rho) = \mathrm{tr}_2\big[U(\rho \otimes |0\rangle\langle 0|)U^\dagger\big]. \tag{10.168}$$

10.2 Applications of the HSW converse

Application 0 is that n qubits can be used to send m classical bits if and only if $m \le n$. So qubits cost a lot but don't give us very much, in line with the old view of quantum mechanics as mostly a nuisance.

Application 1: random access codes (RAC). Suppose $x \in \{0,1\}^m \to \rho_x$ on n qubits. If $m > n$, we cannot reconstruct the entire string x. One might hope to get around this using complementarity: we can measure in one basis or the other, but they are incompatible (measuring in one destroys information in the other). So perhaps, given some i, we can at least reconstruct x_i from ρ_x with high probability (Figure 10.15).

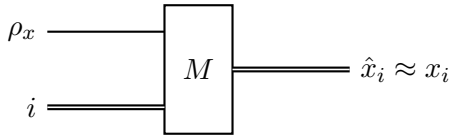


Figure 10.15: Decoding a random access code: given the state ρ_x and an index i (classical, double line), the measurement M outputs a guess $\hat{x}_i$, hopefully equal to x_i.

Let us measure a random access scheme by the worst bit it encodes, meaning we want to find
$$1 - \epsilon = \max_{\text{strategies}} \min_x \min_i \Pr[\hat{x}_i = x_i], \tag{10.169}$$
where the minimum over x can also be replaced by an expectation over x. Classically, we cannot beat $\frac{1}{2}$ if $m > n$ using deterministic encoding. Quantumly, for $m = 2$, $n = 1$, we can achieve[6]
$$\Pr[\hat{x}_i = x_i] = \cos^2(\pi/8) \approx 0.85. \tag{10.170}$$
In general, though, we get at most linear savings:

Theorem 10.2.1 (Nayak)**.** *A random access code encoding m bits into n qubits with error $\epsilon \le 1/2$ must have*
$$n \ge m(1 - H_2(\epsilon)). \tag{10.171}$$

This follows from the Holevo converse, via the following lemma.

[6] One protocol: encode x_1x_2 in the qubit state with Bloch vector $\frac{1}{\sqrt{2}}\big((-1)^{x_2}, 0, (-1)^{x_1}\big)$, i.e. $\cos\alpha\,|0\rangle + \sin\alpha\,|1\rangle$ with $\alpha = (-1)^{x_2}\big(\frac{\pi}{8} + x_1\frac{\pi}{4}\big)$, so the four states are at angles $\pm\frac{\pi}{8}$ and $\pm\frac{3\pi}{8}$. To recover x_1, measure Z; to recover x_2, measure X. Either way, the correct outcome has probability $\frac{1}{2}\big(1 + \frac{1}{\sqrt{2}}\big) = \cos^2(\pi/8)$.

Lemma 10.2.1. *Given σ_0, σ_1, $M_0 + M_1 = I$ such that $\mathrm{tr}[M_b\sigma_b] \geq 1-\epsilon$ for $b \in \{0,1\}$, with $\epsilon \leq 1/2$, let*

$$\omega = \frac{|0\rangle\langle 0|^X \otimes \sigma_0 + |1\rangle\langle 1|^X \otimes \sigma_1}{2}. \tag{10.172}$$

Then $I(X;Q)_\omega \geq 1 - H_2(\epsilon)$.

Proof. Here X is a uniformly random bit, since ω gives σ_0 and σ_1 probability $1/2$ each, so $H(X) = 1$. If we measure Q with $\{M_0, M_1\}$ and get outcome $\hat{X}$, then $\Pr\left[\hat{X} \neq X\right] = \frac{1}{2}(\mathrm{tr}[M_1\sigma_0] + \mathrm{tr}[M_0\sigma_1]) \leq \epsilon$, so by (the proof of) Fano's inequality from Lecture 9 with $d = 2$, $H(X \mid \hat{X}) \leq H_2(\epsilon)$.[7] Therefore

$$I(X;\hat{X}) = H(X) - H(X \mid \hat{X}) \geq 1 - H_2(\epsilon). \tag{10.173}$$

Then

$$1 - H_2(\epsilon) \leq I(X;\hat{X}) \leq I(X;Q), \tag{10.174}$$

where the second inequality comes from data processing. □

Proof of Theorem 10.2.1. Suppose we have a quantum random access code with $x \in \{0,1\}^m \to \rho_x$ on n qubits, with two-outcome measurements $M_0^{(i)} + M_1^{(i)} = I$ indexed by $i \in \{1, \ldots, m\}$, where for all x and i,

$$\mathrm{tr}[M_{x_i}^{(i)}\rho_x] \geq 1 - \epsilon. \tag{10.175}$$

Consider

$$\omega = \frac{1}{2^m} \sum_{x \in \{0,1\}^m} |x\rangle\langle x|^X \otimes \rho_x^Q. \tag{10.176}$$

On one hand, we know $I(X;Q) = S(Q) - S(Q \mid X) \leq n$, since $S(Q) \leq n$ (the entropy is at most the log of the dimension, 2^n) and $S(Q \mid X) \geq 0$ since we are conditioning on something classical ($S(Q \mid X) = 2^{-m}\sum_x S(\rho_x)$). We want to prove

$$I(X;Q) \geq m(1 - H_2(\epsilon)), \tag{10.177}$$

which would give us the desired bound.[8]

By the chain rule, $I(X;Q) = \sum_{i=1}^m I(X_i; Q \mid X_1, \ldots, X_{i-1})$. Note $X_1, \ldots, X_{i-1}$ are classical, and the random access guarantee holds for every x, no matter how we fix these bits. With $x_1, \ldots, x_{i-1}$ fixed, X_i is a uniform bit, and given $X_i = b$, Q is in the state

$$\sigma_b = \frac{1}{2^{m-i}} \sum_{x_{i+1}, \ldots, x_m} \rho_{x_1 \cdots x_{i-1} b\, x_{i+1} \cdots x_m}. \tag{10.178}$$

Since $\mathrm{tr}[M_b^{(i)}\rho_x] \geq 1 - \epsilon$ for every ρ_x in this average, $\mathrm{tr}[M_b^{(i)}\sigma_b] \geq 1 - \epsilon$. So by Lemma 10.2.1, each one of these mutual informations is at least $1 - H_2(\epsilon)$, and we get $I(X;Q) \geq m(1 - H_2(\epsilon))$. □

This means we cannot compress a string by more than a constant factor, even if we only want to be able to access a single bit. Classically, we could not beat $1/2$ only because we forced the encoding to be deterministic. With random choices that the adversary picking x and i cannot see,

[7] This is where the proof breaks down for $\epsilon > 1/2$. The binary entropy H_2 is increasing only on $[0, 1/2]$, so $\Pr\left[\hat{X} \neq X\right] \leq \epsilon$ implies $H_2(\Pr\left[\hat{X} \neq X\right]) \leq H_2(\epsilon)$ only when $\epsilon \leq 1/2$. For $\epsilon > 1/2$, Theorem 10.2.1 is false: guessing each x_i at random, with no qubits at all, already has error $1/2 < \epsilon$, while $m(1 - H_2(\epsilon)) > 0$.

[8] $1 - H_2(\epsilon)$ is the capacity of the binary symmetric channel with flip probability ϵ.

we can do better. For example, to encode 2 million bits into 1 million and 1, we encode either the first half or the last half (chosen at random), and the final bit encodes which one we did. For any x and i, the decoder knows x_i with probability $1/2$ and otherwise guesses, giving a success rate of $3/4$. With randomized encoding, the same bound $n \geq m(1 - H_2(\epsilon))$ is achievable classically up to a sublinear correction, so Theorem 10.2.1 is tight and we are barely getting a quantum advantage.

10.3 Quantum state learning

Application 1.5 applies the random access bound to learning quantum states. In sensing (Lecture 7), we learn one parameter; in state tomography, we learn all of them, outputting $\hat{\rho} \approx \rho$ from many copies of ρ. Shadow tomography is in between: given a family of measurements $\mathcal{M}$, we output $\hat{\rho}$ such that for all $M \in \mathcal{M}$,

$$\left|\mathrm{tr}[M(\rho - \hat{\rho})]\right| \leq \text{small}. \tag{10.179}$$

Here we think of M as the two-outcome measurement $\{M, I - M\}$. Another version of this is called PAC learning (probably approximately correct), where we have a distribution D over measurements, and we want

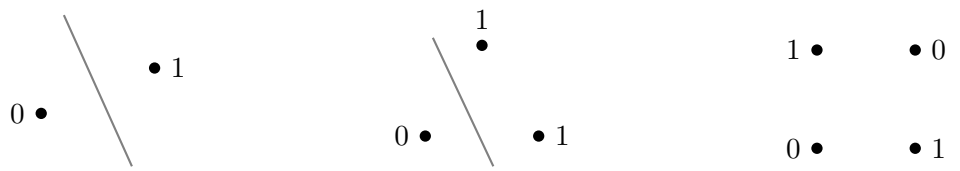
$$\mathbb{E}_{M \sim D}\left|\mathrm{tr}[M(\rho - \hat{\rho})]\right| \leq \epsilon. \tag{10.180}$$

It turns out PAC learning is way cheaper than state tomography. For state tomography of a d-dimensional state we need $\gtrsim d^2$ copies (Section 10.4), while for PAC learning $O(\log d)$ samples suffice (on average over $M \sim D$, regardless of what the family $\mathcal{M}$ is). Bounds like these tell us whether the difficulty of a learning task is information-theoretic or computational.

The proof of the PAC learning bound is not self-contained. Consider $\rho : \mathcal{M} \to [0, 1]$ mapping $M \mapsto \mathrm{tr}[M\rho]$, and assume $\mathrm{tr}[M\rho] \geq 1 - \epsilon$ or $\leq \epsilon$. This is not necessary, but it simplifies things, since then ρ basically just labels each measurement 0 or 1, and we can use the classical results on learning labelings.

Definition 10.3.1 (VC-dimension)**.** *A set of points $\{M_1, \ldots, M_d\}$ is shattered by a set S of $\{0,1\}$-valued functions if for every labeling $o \in \{0,1\}^d$ there is an $f \in S$ with $f(M_j) = o_j$ for all j. For a set S, such as the set of n-qubit states (under the assumption above), the VC-dimension VC-dim(S) (after Vapnik and Chervonenkis) is the size of the largest set shattered by S.*

For a classical example, we label points in the plane 0 or 1 using $S = \{\text{half-spaces}\}$: we draw a line, and label the points on one side 0 and the points on the other side 1. In two dimensions, two points can be shattered: the line goes between them exactly when their labels differ (Figure 10.16). Three non-collinear points can be shattered as well. Four points cannot, as shown in Figure 10.16(c). So the VC-dimension of half-spaces in the plane is 3, or in general dim +1.



(a) 2 points (b) 3 points (c) 4 points

Figure 10.16: Shattering by half-spaces in the plane. (a), (b) For 2 or 3 points, every labeling can be realized by some line; one labeling is shown for each. (c) For 4 points in a square with the diagonals labeled 0 and 1, no line separates the 0s from the 1s.

A theorem that we will not prove is that the number of examples needed to PAC learn a set S is $\lesssim$ VC-dim(S). Here there is an unknown element $f \in S$ and an unknown distribution over inputs; we train on examples $(M, f(M))$ with M drawn from this distribution, and are tested on fresh inputs from the same distribution.

For n qubits, $\mathrm{tr}[M\rho]$ is a number in $[0,1]$ rather than a label in $\{0,1\}$, so we actually want the fat-shattering dimension.

Definition 10.3.2 (Fat-shattering dimension)**.** *The fat-shattering dimension of n-qubit states is the maximum d for which there exist $M_1, \ldots, M_d \in \mathcal{M}$ such that for every labeling $o \in \{0,1\}^d$ there is a state ρ such that for all j,*

$$\left|\mathrm{tr}[M_j\rho] - o_j\right| \leq \epsilon. \tag{10.181}$$

This measures the richness of the class of n-qubit states: a class that can fit any pattern of outcomes on many measurements is too rich to learn from few examples.

By the RAC no-go theorem,

$$d \leq \frac{n}{1 - H_2(\epsilon)}, \tag{10.182}$$

since fat-shattering measurements give a random access code for d bits. Given fat-shattering measurements $M_1, \ldots, M_d$, say we want to encode $x \in \{0,1\}^d$; fat-shattering guarantees a ρ_x with $|\,\mathrm{tr}[M_j\rho_x] - x_j| \leq \epsilon$ for all j. To extract x_i, note that measuring $\{M_i, I - M_i\}$ gives outcome M_i with probability at least $1-\epsilon$ if $x_i = 1$ and at most ϵ if $x_i = 0$, so outputting 1 on outcome M_i gives x_i with probability at least $1-\epsilon$. So we have a random access code with error ϵ encoding d bits into n qubits, and Theorem 10.2.1 gives $n \geq d(1 - H_2(\epsilon))$.

As a corollary, we can PAC learn n-qubit states with $O(n)$ examples (pairs $(M, \mathrm{tr}[M\rho])$). Here each example gives the exact value $\mathrm{tr}[M\rho]$. With some more work, we can make it work with examples that only give a measurement outcome, i.e. a sample from the outcome distribution $(\mathrm{tr}[M\rho], 1 - \mathrm{tr}[M\rho])$.

The PAC learning result is also related to multiplicative weights ("go with the winner") and to quantum algorithms for solving large SDPs.

10.4 State tomography

Application 2 is state tomography. Given $\rho^{\otimes n}$, where ρ is a $d \times d$ density matrix, output $\hat{\rho}$ such that

$$\mathbb{E}_{\hat{\rho}} \frac{1}{2}\|\rho - \hat{\rho}\|_1 \leq \epsilon. \tag{10.183}$$

Recall $\frac{1}{2}\|\rho - \hat{\rho}\|_1$ is the trace distance, and the expectation is over the randomness of the measurement outcomes, since $\hat{\rho}$ is a random variable.

Theorem 10.4.1. *Any such procedure needs $n \gtrsim d^2/\epsilon^2$ copies.*

Proof idea. Suppose a procedure uses n copies. We pick test states $\rho_1, \ldots, \rho_m$, all within $O(\epsilon)$ of I/d, with $\|\rho_x - \rho_y\|_1 \geq 4\epsilon$ for all $x \neq y$; one can show $m = \exp\left(cd^2\right)$ such states exist, for a constant $c > 0$. Pick x at random and give the procedure $\rho_x^{\otimes n}$. Outputting the $\hat{x}$ with $\rho_{\hat{x}}$ closest to $\hat{\rho}$ recovers x with probability $\geq 1 - \delta$ (by the triangle inequality, adjusting constants via Markov's inequality). With

$$\omega = \frac{1}{m}\sum_{x=1}^{m} |x\rangle\langle x|^X \otimes \rho_x^{\otimes n}, \tag{10.184}$$

Fano's inequality and data processing give

$$I(X;Q^n) \geq (1-\delta)\log m - 1 \gtrsim d^2. \tag{10.185}$$

On the other hand, by subadditivity as in the HSW converse,

$$I(X;Q^n) \leq n\, I(X;Q_1) \leq n\big(\log d - \mathbb{E}_x S(\rho_x)\big) \lesssim n\epsilon^2, \tag{10.186}$$

since $S(\rho_x) \approx \log d - O(\epsilon^2)$. So $d^2 \lesssim n\epsilon^2$. □

8.372 Quantum Information Science III **Fall 2026**

Lecture 11: October 20, 2026

Scribe: Aarush Vailaya *State tomography lower bound and resource inequalities*

11.1 State tomography lower bound

We finish the lower bound $n \gtrsim d^2/\epsilon^2$ on state tomography from Lecture 10. Recall the task: given $\rho^{\otimes n}$, where ρ is a $d \times d$ density matrix, output $\hat{\rho}$ with $\mathbb{E}_{\hat{\rho}} \frac{1}{2}\|\rho - \hat{\rho}\|_1 \leq \epsilon$. The idea of the lower bound is to give the procedure one of m well-separated test states near I/d. A good tomography procedure would identify which state it got, which is $\log m \gtrsim d^2$ bits of information, but each copy of a nearly maximally mixed state carries only $O(\epsilon^2)$ bits (meaning $n \gtrsim d^2/\epsilon^2$ copies are necessary).

Consider an ϵ-packing net of $\exp\big(\Omega(d^2)\big)$ states with entropy $\geq \log d - O(\epsilon^2)$. That is, we want states $\rho_1, \ldots, \rho_m$ such that

1. **(ϵ-packing)** $\|\rho_x - \rho_y\|_1 > \epsilon$ for all $x \neq y$,
2. **(many states)** $m = \exp\big(\Omega(d^2)\big)$,
3. **(high entropy)** $S(\rho_x) \geq \log d - O(\epsilon^2)$ for all x.

An ϵ-packing is a set of points whose pairwise distances are all greater than ϵ, while an ϵ-covering of a set S is a set of points such that every point of S is within ϵ of one of them. In the set of all states, the states with $S(\rho) \geq \log d - O(\epsilon^2)$ form a small ball with I/d at its center, and the states of the net (the packing) are spread out inside it (Figure 11.17).

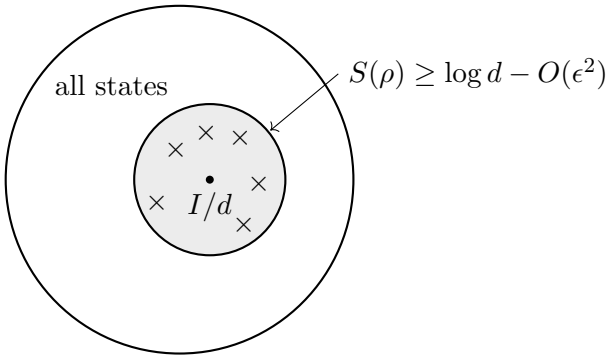


Figure 11.17: The states with entropy $\geq \log d - O(\epsilon^2)$ form a small ball around I/d inside the set of all states. The states ρ_x of the packing net ($\times$) lie in this ball and are pairwise far apart.

Consider the state

$$\omega = \frac{1}{m}\sum_{x=1}^{m} |x\rangle\langle x|^X \otimes \rho_x^{\otimes n}, \tag{11.187}$$

where X is the label $|x\rangle\langle x|$ and $Q^n = Q_1 \cdots Q_n$ holds $\rho_x^{\otimes n}$, with $\hat{X}$ the decoded label. Suppose we have tomography that can estimate your state up to ϵ, with failure probability δ. That is enough to distinguish these states (the $\times$ marks in Figure 11.17), if $\hat{x}$ is obtained by measuring $\rho_x^{\otimes n}$ with the tomography protocol and choosing the $\rho_{\hat{x}}$ closest to its estimate: with n copies we

are able to measure them and get an ϵ-precise estimate, and then $\hat{x} = x$ with probability at least $1 - \delta$. (For this, the precision of the tomography should be a constant factor smaller than the separation in condition 1; we ignore such constant factors.) That is what an ϵ-packing net and a good tomography protocol give: by Fano's inequality,[9] condition 1 gives

$$I(X;\hat{X}) \geq S(X)(1-\delta) - 1. \tag{11.188}$$

The second condition gives $S(X) = \log m \gtrsim d^2$, which is what we want, since a $d \times d$ density matrix has $d^2 - 1$ real parameters, and this is d^2 bits. Even though a real number is worth more than a bit, once we take precision into account, each parameter is worth only a constant number of bits.

Finally, the third (entropy) condition means $I(X;Q^n) \lesssim n\epsilon^2$, since by subadditivity $I(X;Q^n) \leq n\, I(X;Q_1)$,[10] and the mutual information $I(X;Q_1)$ is the entropy of the average state, which is at most $\log d$, minus the average of the entropies, which is at least $\log d - O(\epsilon^2)$:

$$I(X;Q^n) \leq n\, I(X;Q_1) = n\Big[S\Big(\frac{1}{m}\sum_x \rho_x\Big) - \frac{1}{m}\sum_x S(\rho_x)\Big] \leq n\big[\log d - (\log d - O(\epsilon^2))\big] = O(n\epsilon^2). \tag{11.189}$$

This is an application of data processing, since we have a Markov chain $X - Q^n - \hat{X}$, so $I(X;\hat{X}) \leq I(X;Q^n)$. Combining this with (11.188) and (11.189),

$$d^2 \lesssim S(X)(1-\delta) - 1 \leq I(X;\hat{X}) \leq I(X;Q^n) \lesssim n\epsilon^2, \tag{11.190}$$

meaning $n \gtrsim \frac{d^2}{\epsilon^2}$. This is the whole argument here, except that we need to construct the states with these properties: they need to be close to maximally mixed (condition 3), there needs to be a lot of them (condition 2), and they need to be far apart (condition 1).

11.2 Constructing the states

Let us first construct one state with the correct entropy, starting with a single qubit. Let

$$\rho = \frac{1+\epsilon}{2}\,|0\rangle\langle 0| + \frac{1-\epsilon}{2}\,|1\rangle\langle 1|\,, \tag{11.191}$$

then $S(\rho) = H_2\big(\frac{1+\epsilon}{2}\big)$ (Figure 11.18). Since $\epsilon = 0$ is a local max, there can be no $O(\epsilon)$ terms, so we have

$$H_2\Big(\frac{1+\epsilon}{2}\Big) = 1 - O(\epsilon^2) + O(\epsilon^4) \tag{11.192}$$

(calculus gives the coefficients).

So consider (for even d)

$$\rho \otimes \frac{I_{d/2}}{d/2}, \qquad S\Big(\rho \otimes \frac{I_{d/2}}{d/2}\Big) = H_2\Big(\frac{1+\epsilon}{2}\Big) + \log\frac{d}{2} = \log d - O(\epsilon^2). \tag{11.193}$$

[9] Fano's inequality (Lecture 9) gives $H(X \mid \hat{X}) \leq 1 + \delta \log m$. Since X is uniform on m values, $S(X) = \log m$, so $I(X;\hat{X}) = S(X) - H(X \mid \hat{X}) \geq S(X)(1-\delta) - 1$.

[10] Here $S(Q^n \mid X) = \frac{1}{m}\sum_x S(\rho_x^{\otimes n}) = n \cdot \frac{1}{m}\sum_x S(\rho_x)$ exactly, since entropy is additive on $\rho_x^{\otimes n}$. Subadditivity (Lecture 9) gives $S(Q^n) \leq \sum_{i=1}^n S(Q_i) = n\, S(\bar{\rho})$, where $\bar{\rho} = \frac{1}{m}\sum_x \rho_x$. Subtracting gives $I(X;Q^n) = S(Q^n) - S(Q^n \mid X) \leq n\, I(X;Q_1)$.

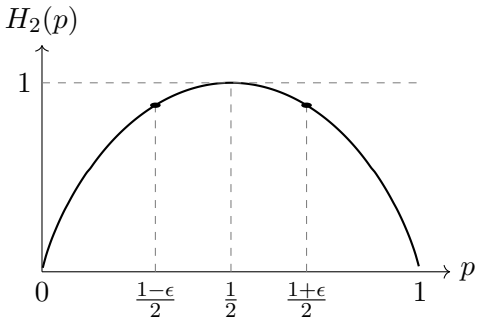


Figure 11.18: The binary entropy $H_2(p)$. Its maximum at $p = \frac{1}{2}$ is quadratic, so moving to $p = \frac{1\pm\epsilon}{2}$ only lowers the entropy by $O(\epsilon^2)$.

This is condition 3. Its eigenvalues are $\frac{1+\epsilon}{d}$ ($d/2$ times) and $\frac{1-\epsilon}{d}$ ($d/2$ times). This is one such state, but we want to construct many such states. To do this, consider using a unitary

$$\rho_x = U_x\Big(\rho \otimes \frac{I_{d/2}}{d/2}\Big)U_x^\dagger, \tag{11.194}$$

which does not change the entropy since unitaries preserve the eigenvalues.

To get an ϵ-packing net we must choose the unitaries, and the question becomes how to choose the unitaries so that the ρ_x are well separated. Geometrically, conditions 1 and 2 are about the trace norm (1-norm), but condition 3 is really about the 2-norm: for ρ near I/d, $\log d - S(\rho) \approx \frac{d}{2\ln 2}\big\|\rho - \frac{I}{d}\big\|_2^2$.

To choose the unitaries, first recall how to construct packing nets in general. Suppose we have a general norm $\|\cdot\|$ in a space of dimension D and let $B = \{x : \|x\| \le 1\}$. Then $\mathrm{vol}(\epsilon B) = \epsilon^D \mathrm{vol}(B) = \epsilon^D$ (normalizing the volume of the unit ball to be 1). We can use a greedy algorithm, where we take any point in the ball and draw a ball of radius ϵ around it, then take another point outside that ball and draw a ball of radius ϵ around it, and keep proceeding (Figure 11.19). At a certain point we cannot add any more points, because the entire big ball is shaded. The points are pairwise more than ϵ apart, and each ball covers at most ϵ^D volume, so there exists a packing net of size at least $\frac{1}{\epsilon^D}$. A similar argument says there exists a covering net of size at most $(1 + 2/\epsilon)^D$. This argument works if the set we are trying to cover is the unit ball of a norm, and the ϵ-balls are in the same norm. We will use this later in the class.

For our case, we will instead use the following lemma. The construction will be to choose the U_x randomly, and argue that two random states are unlikely to be close:

Lemma 11.2.1. *For independent random unitaries U_x, U_y and ρ_x, ρ_y as in (11.194),*

$$\Pr_{U_x,U_y}\big(\|\rho_x - \rho_y\|_1 \le \epsilon\big) \lesssim \exp\big(-\Omega(d^2)\big). \tag{11.195}$$

Here the ϵ in the lemma matches the ϵ in (11.194) up to constant factors. Assuming the lemma is true, we can construct the net as follows: given a net of size m made of random states, the union bound says

$$\Pr(\text{net not } \epsilon\text{-packing}) \le \binom{m}{2} e^{-\Omega(d^2)}, \tag{11.196}$$

so we can choose m of order e^{cd^2} (for a small enough constant $c > 0$) and still have this probability below 1, and the probabilistic method implies there is a net that doesn't fail. This will not be

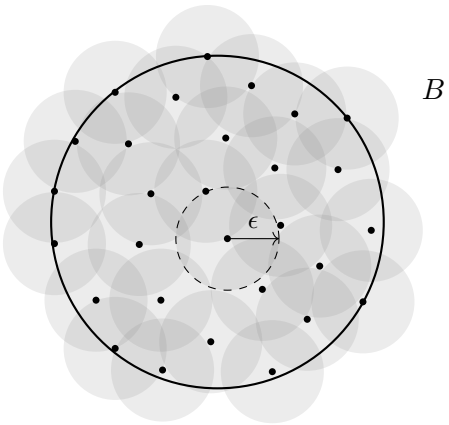


Figure 11.19: Greedy construction of a packing net in two dimensions. Each new point is chosen outside the ϵ-balls (shaded) around the previous points, so the points are pairwise more than ϵ apart. The process stops when the shaded balls cover the unit ball B.

proven today, but the next theme of the class will be random objects in quantum information and we will build towards proving this.

We could apply this to pure states, and we would get $\exp(-\Omega(d))$ instead of $\exp\left(-\Omega(d^2)\right)$ (since pure states only have $O(d)$ parameters). Lemma 11.2.1 is similar to the Chernoff bound, which says that if we have many degrees of freedom, the probability that their average deviates from its expectation is exponentially small.

Note that we are not packing the whole ball of high-entropy states, but only the spherical shell of states with eigenvalues $\frac{1\pm\epsilon}{d}$, since we are restricting ourselves to one unitary orbit. But this is not a restrictive thing to do, because the resulting lower bound is tight: tomography is possible with $O(d^2/\epsilon^2)$ copies. Additionally, the states that are very mixed make up most of the volume, which is why we don't need to consider every possible state.

We started off looking at Shannon's theorems, and Shannon's noisy coding theorem used a random codebook. Instead of trying to find a structured code, he showed that a good codebook exists using the probabilistic method. We want to do something similar in the quantum case, so we are going to study random quantum states, random unitaries, and analogues of what Shannon did, starting on Thursday.

11.3 Information theory beyond classical communication

So far, we have seen how to use a quantum channel to send classical bits. However, with a quantum channel, we may want to send quantum bits. Let us jump beyond that to see in general what we may want to do with quantum channels.

- We may have resources like a channel $\mathcal{N}^{A\to B}$ or a state ρ^{AB},
- and goals such as cbits, qubits, ebits, pbits (private / secret bits), and rbits (shared randomness).
- We allow assistance / auxiliary resources such as classical communication (one-way or two-way) and shared entanglement.

A lot of channel protocols can be viewed as resource inequalities, such as

$$\langle \mathcal{N} \rangle \geq C(\mathcal{N}) \text{ cbits}, \tag{11.197}$$
$$\langle \mathcal{N} \rangle + \infty \text{ ebits} \geq C_E(\mathcal{N}) \text{ cbits}, \tag{11.198}$$
$$\langle \mathcal{N} \rangle + \infty \text{ ebits} \geq Q_E(\mathcal{N}) \text{ qubits}, \tag{11.199}$$
$$\langle \mathcal{N} \rangle \geq Q(\mathcal{N}) \text{ qubits}, \tag{11.200}$$
$$\langle \rho \rangle + \infty \text{ cbits}_{\rightarrow} \geq E_D(\rho) \text{ ebits}. \tag{11.201}$$

In (11.197), $\langle \mathcal{N} \rangle$ is the resource of being able to use the channel and $C(\mathcal{N})$ is the channel capacity. So using the channel, we can send that many cbits, but the $\geq$ hides complexity like the fact that this is asymptotic and approximate. Similarly, C_E and Q_E are the entanglement-assisted classical and quantum capacities, and Q is the quantum capacity. The last inequality (11.201) is called entanglement distillation. Consider noisy entanglement between two parts of a computer, from which we want to distill asymptotically pure EPR pairs. Typically we allow classical communication (here one-way, from Alice to Bob), and the optimal rate is called the distillable entanglement, $E_D(\rho)$.

Beyond the non-trivial ones, there are basic resource inequalities like

$$1 \text{ qubit} \geq 1 \text{ cbit} \geq 1 \text{ rbit}, \qquad 1 \text{ qubit} \geq 1 \text{ ebit} \geq 1 \text{ rbit}, \tag{11.202}$$

and slightly more nontrivial ones like

$$2 \text{ cbits} + 1 \text{ ebit} \geq 1 \text{ qubit} \qquad \text{(teleportation)}, \tag{11.203}$$
$$1 \text{ qubit} + 1 \text{ ebit} \geq 2 \text{ cbits} \qquad \text{(superdense coding)}. \tag{11.204}$$

We don't want to go too much into these quantities because the existing state of knowledge is not that good. We have seen a formula for $C(\mathcal{N})$ and $C_E(\mathcal{N})$,[11] and later we will see a formula for Q (but unfortunately, it isn't nearly as nice as the C formula or the nicer C_E formula). Even the question of when $Q(\mathcal{N}) = 0$ is hard.

Going back to classical capacity, a quantum channel has zero classical capacity when it is a constant channel. Conversely, if it is not a constant channel, then Alice can produce at least two distinguishable states for Bob, which can be used to send a bit with enough repetition. But for the quantum case, this is very hard, which we will see by looking at the analogous questions for states. Here, we have a state ρ^{AB} with distillable entanglement $E_D(\rho)$, and denote $E_{D,2}(\rho)$ as its distillable entanglement with two-way communication. Going the other way, the entanglement cost $E_C(\rho)$ is how much entanglement is needed to produce ρ from EPR pairs: it is the minimum E_C such that

$$E_C \text{ ebits} + \infty \text{ cbits} \geq \langle \rho \rangle, \tag{11.205}$$

where $\langle \rho \rangle$ is again the resource corresponding to the state ρ. Unlike distillable entanglement, this is the cost to produce the state, which sometimes differs from the distillable entanglement.

These questions are inspired by how well things work for pure states. Consider $|\Psi\rangle^{AB}$. Then

$$\langle \Psi \rangle =_{\text{LOCC}} E(\Psi) \text{ ebits}, \qquad E(\Psi) = S(\Psi^A), \tag{11.206}$$

where $E(\Psi)$ is called the entropy of entanglement. Going from many copies of Ψ to EPR pairs is called entanglement concentration, and the other way is called entanglement dilution. These are

[11] From Lecture 10, $C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n}\chi(\mathcal{N}^{\otimes n})$. The entanglement-assisted capacity is $C_E(\mathcal{N}) = \max_{\phi^{RA}} I(R;B)_{(\mathrm{id}_R \otimes \mathcal{N})(\phi)}$, maximized over pure states ϕ^{RA} and with no limit over n.

asymptotically reversible, which made people hopeful that for mixed states we can get something nice.

For mixed states though, in general all we can say is $E_C \geq E_D$, and sometimes the inequality is strict. There is also the issue of formulas. For pure states, $E(\Psi)$ is easy to calculate, but for mixed states we need to define E_C and E_D in terms of asymptotic operations. There is another quantity called the entanglement of formation.

Definition 11.3.1 (Entanglement of formation)**.**

$$E_F(\rho) = \min_{\substack{\{p_i,\psi_i\} \\ \rho=\sum_i p_i\psi_i}} \sum_i p_i E(\psi_i). \tag{11.207}$$

From this picture it is clear this gives one way to prepare ρ (prepare each ψ_i with probability p_i, using $E(\psi_i)$ ebits per copy), so $E_F \geq E_C \geq E_D$. This inequality is also sometimes strict, and in fact

$$E_C(\rho) = \lim_{n\to\infty} \frac{1}{n} E_F(\rho^{\otimes n}), \tag{11.208}$$

and the size of the effect of two vs. one copies is in general unknown. There is a similar formula for distillable entanglement:

$$E_D(\rho) = \lim_{n\to\infty} \frac{1}{n} \max_{\mathcal{E}:A^n\to A'} \big[S(B^n) - S(A'B^n)\big]. \tag{11.209}$$

Like Shannon's theorem, there are entropic quantities that we are maximizing, but unlike Shannon's theorem we have an increasing number of copies. Moreover, in Shannon's theorem we are maximizing a concave function over a convex set, which makes it efficient (in some sense) to compute, but here we do not have that property. What is good about it is that it is good as a lower bound (each fixed n gives a lower bound on E_D), but the converse is difficult.

11.4 Separability and bound entanglement

One thing we can say a lot more about is the zero condition, like $E_F(\rho) = 0$. This means

$$\rho = \sum_i p_i \, \alpha_i \otimes \beta_i \tag{11.210}$$

is separable, since there is a decomposition $\rho = \sum_i p_i\psi_i$ in which each $E(\psi_i) = S(\psi_i^A) = 0$, so each $\psi_i = \alpha_i \otimes \beta_i$ is a product state. One difference is that a product state has no correlation, while a separable state can have classical correlations. We say a mixed state is entangled if and only if it is not separable. While this is pretty easy, something very non-trivial is that

$$E_C(\rho) = 0 \iff E_F(\rho) = 0. \tag{11.211}$$

One direction is trivial (since $E_C \leq E_F$), but the other direction is not: if we could prepare our entangled state using only $n^{0.9}$ ebits for n copies, then $E_C(\rho) = 0$ even though the state is entangled, but these states don't exist. This will not be proven in class.

One unpleasant thing is that figuring out whether ρ is separable is NP-complete.[12] The interesting thing about this question (which Prof. Harrow has researched) is that we are glossing over

[12]NP is the class of decision problems whose "yes" answers can be verified efficiently given a proof. A problem is NP-hard if every problem in NP can be efficiently reduced to it, and NP-complete if it is both NP-hard and in NP. The easy direction is to show that separability is in NP, since given a decomposition $\rho = \sum_i p_i\alpha_i \otimes \beta_i$ we can easily verify that it adds up to ρ. We did not show why it is NP-hard. NP-hardness means that (assuming P $\neq$ NP) there is no efficient algorithm for deciding separability, and in particular no simple formula for it.

the approximation question. In particular, NP is always referring to some discrete problem, but these have real numbers. We can ask in general the weak membership problem, where we promise that we are either inside the set or ϵ-far from the set (Figure 11.20). In particular, the algorithm must be right if ρ is separable or pretty far away from separable, and there are no guarantees in the intermediate regime. The complexity depends on ϵ, and if $\epsilon \approx \frac{1}{\text{poly(dimension)}}$, then we get NP-complete. However, if ϵ is constant, it seems the complexity goes as quasipolynomial, $d^{\log d}$. We will go over some of the arguments for that in a few weeks.

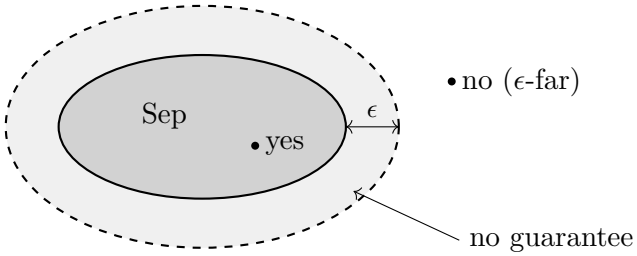


Figure 11.20: The weak membership problem for Sep. The algorithm must answer "yes" on separable states and "no" on states ϵ-far from Sep (outside the dashed line). In the band in between, either answer is allowed.

Something we understand less well is when $E_D(\rho) = 0$. We just know a few sufficient (but not necessary) conditions, but the sufficient conditions are pretty useful. If $E_D(\rho) = 0$ and $\rho \notin \text{Sep}$, we call this "bound entanglement". Bound entangled states are like the black holes of entanglement: it takes entanglement to create them, but you cannot get any useful entanglement out. We don't have a complete characterization of these.

2-extendibility. One sufficient condition comes from no-cloning or monogamy of entanglement.

Definition 11.4.1 (2-extendible)**.** *ρ^{AB} is 2-extendible if $\exists\, \tilde{\rho}^{AB_1B_2}$ such that $\rho^{AB} = \tilde{\rho}^{AB_1} = \tilde{\rho}^{AB_2}$ (Figure 11.21).*

It is important that we simply ask whether there exists a state with these properties.

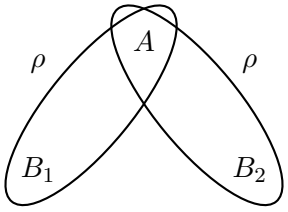


Figure 11.21: A 2-extension $\tilde{\rho}^{AB_1B_2}$ of ρ^{AB}: both of the reduced states on AB_1 and on AB_2 equal ρ.

One example of a state that is not 2-extendible is any pure entangled state $|\Psi\rangle^{AB}$. That is because a pure state cannot be correlated with anything else: if $\tilde{\rho}^{AB_1} = \Psi$, then $\tilde{\rho}^{AB_1B_2} = \Psi^{AB_1} \otimes \sigma^{B_2}$, so $\tilde{\rho}^{AB_2} = \Psi^A \otimes \sigma^{B_2}$, which is a product state and hence not Ψ. This is sometimes called the principle of "monogamy of entanglement", which distinguishes it from classical shared randomness. If we have a definite EPR pair, we know that our qubit is entangled with the other qubit and nothing else.

This shows that 2-extendible states are not distillable with one-way communication, which means communication from Alice to Bob. Suppose for contradiction that we could distill something

from a 2-extendible ρ^{AB} using one-way communication. Then apply the same protocol to n copies of $\tilde{\rho}^{AB_1B_2}$: Alice broadcasts her messages to both B_1 and B_2 (classical messages can be copied), and both Bobs follow Bob's part of the protocol. Since $\tilde{\rho}^{AB_1} = \tilde{\rho}^{AB_2} = \rho^{AB}$, each one will correctly distill some approximately pure entangled states, giving us approximate EPR pairs between Alice's output and B_1, and also between Alice's output and B_2, meaning there isn't enough of Alice to go around (Alice is using the same register for both). This argument breaks with two-way communication, since then B_1 and B_2 would each send their own messages back to Alice, and she cannot respond to both.

PPT. The second condition is that ρ is PPT (positive under partial transpose). Write $(\mathrm{id} \otimes T)(\rho) = \rho^\Gamma$ (since Γ is half of a T), where T is the transpose on B, so

$$(|i\rangle\langle j| \otimes |k\rangle\langle l|)^\Gamma = |i\rangle\langle j| \otimes |l\rangle\langle k|. \tag{11.212}$$

The PPT states are $\{\rho : \rho^\Gamma \geq 0\} \supsetneq \mathrm{Sep}$ (separable states are PPT since $(\alpha \otimes \beta)^\Gamma = \alpha \otimes \beta^T \geq 0$), and these are contained in the set of states with zero distillable entanglement (even with two-way communication). This is a quick argument, which we don't do in detail now: local operations and classical communication preserve PPT, so distillation won't change the fact that you are PPT, but pure entangled states are not PPT.

Really quick, let us mention the channel versions of these. The channel version of separable is entanglement-breaking or QCQ (Lecture 10). The channel version of 2-extendible is an antidegradable channel. An example is the erasure channel with erasure probability $\geq 50\%$. A quantum erasure channel with probability $1/2$ erases the qubit half the time (meaning Eve gets the original qubit), so half the time it goes to Bob and half the time it goes to Eve, so we cannot distill entanglement with Bob (unless Bob can send messages back) as otherwise Eve could make the same entanglement. This is antidegradable since if Eve adds noise (erasing some of her qubits), she can get the same view as Bob. The channel version of PPT is PPT channels (channels whose Choi state is PPT). These are all channels with zero quantum capacity.

8.372 Quantum Information Science III **Fall 2026**

Lecture 12: October 22, 2026

Scribe: Aarush Vailaya *Randomness and concentration of measure*

12.1 Scalar random variables

In classical information theory, things like random codebooks were useful, and we want quantum analogues of them. We will talk about scalar random variables, vector random variables, and matrix random variables. We want to study measure concentration, which is showing that a random variable is close to its mean.

The first inequality is Markov's inequality, which states that if $X \geq 0$ then

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{12.213}$$

For instance, if the average height in the room is five feet, no more than half of the people can have a height of 10 feet or more, since otherwise the average would be too high (since no one can have negative height). In particular, this bound isn't too strong. To prove it, write

$$\mathbb{E}[X] = \int_0^\infty dt \, \Pr[X \geq t]. \tag{12.214}$$

The integrand looks like a monotonically decreasing function of t (Figure 12.22), and the integral is the area under it. We do an inner approximation by the rectangle of width a and height $\Pr[X \geq a]$, giving $\mathbb{E}[X] \geq a \Pr[X \geq a]$.

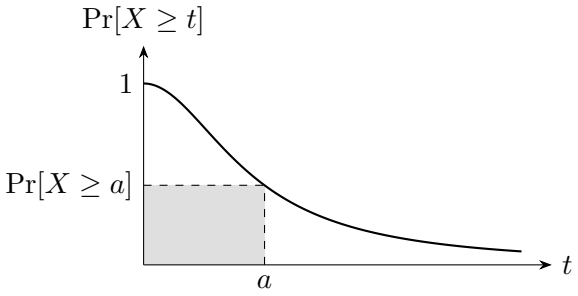


Figure 12.22: Proof of Markov's inequality. $\mathbb{E}[X]$ is the area under the decreasing curve $t \mapsto \Pr[X \geq t]$. The shaded rectangle of width a and height $\Pr[X \geq a]$ lies under the curve, so $\mathbb{E}[X] \geq a \Pr[X \geq a]$.

To get stronger bounds, we use more information. Using the standard deviation σ of X, we get Chebyshev's inequality, which states

$$\Pr\big[|X - \mathbb{E}[X]| \geq a\sigma\big] \leq \frac{1}{a^2}. \tag{12.215}$$

The proof is to use $Y = (X - \mathbb{E}[X])^2$ (note $\mathbb{E}[Y] = \sigma^2$) and apply Markov to this:

$$\Pr\big[|X - \mathbb{E}[X]| \geq a\sigma\big] = \Pr\big[Y \geq a^2\sigma^2\big] \leq \frac{\mathbb{E}[Y]}{a^2\sigma^2} = \frac{1}{a^2}.$$

One way to think about this: suppose I give someone \$5 and penalize them by whatever number they send back, which incentivizes them to give the number 0. But something better I can do is give \$25, and penalize them by the square of the number they send. Both deals break even at 5, but the second punishes large numbers much more, just as Chebyshev (Markov applied to $(X-\mathbb{E}[X])^2$) beats Markov applied to $|X-\mathbb{E}[X]|$. So one might ask why not use even higher moments, the expectations $\mathbb{E}[X^k]$ for larger k.

But the issue is that higher moments might diverge. Professor Harrow mentioned a project he worked on as an undergraduate. He was very proud of it until he refreshed his browser and saw the one blue link showing it had been solved in the 70s. He was trying to calculate the variance of a signal with $1/t^2$ tails, and the issue was that the variance actually grew with the size of the measurement. More generally, say $\Pr[X=x] \sim x^{-\alpha}$. This requires $\alpha > 1$ (assuming X can go on to infinity) for normalization, but for $\mathbb{E}[X] < \infty$, we need $\alpha > 2$ (since $\mathbb{E}[X] \sim \int^\infty dx\, x^{1-\alpha}$). For $\operatorname{Var} X < \infty$, we need $\alpha > 3$. There are perfectly reasonable random variables where the first moment is defined but the second moment is not defined.

Higher moments are also harder to calculate. Consider the moment generating function $\mathbb{E}[e^{\lambda X}] = \sum_n \frac{\lambda^n}{n!}\mathbb{E}[X^n]$. Here $\mathbb{E}[X^n]$ is called the n-th moment of X, and the moment generating function gets its name because its Taylor coefficients are exactly these moments, packaged into a single function of λ. If $\lambda > 0$ then $x \mapsto e^{\lambda x}$ is an increasing function. Thus,

$$\Pr[X \geq a] = \Pr\left[e^{\lambda X} \geq e^{\lambda a}\right] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda a}}. \tag{12.216}$$

Like Chebyshev, this is Markov's inequality applied to a function of X, here $Y = e^{\lambda X}$: X is large exactly when Y is large. This is useful if $X = X_1 + \cdots + X_n$ is the sum of independent random variables $\{X_i\}$.[13] Then, using independence,

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}]. \tag{12.217}$$

In terms of moments, this works because independence turns every mixed moment (of the form $\mathbb{E}[X_1^{k_1}\cdots X_n^{k_n}]$) into a product of moments of the individual variables, $\prod_i \mathbb{E}[X_i^{k_i}]$. As a simple example, say each $X_i = \pm 1$, uniformly at random.[14] Then,

$$\mathbb{E}[e^{\lambda X_i}] = \frac{e^{-\lambda} + e^{\lambda}}{2} = \cosh\lambda \leq e^{\lambda^2/2}. \tag{12.218}$$

It is good to check the last inequality with pen and paper once.[15] Note $\operatorname{Var} X_i = 1$, so X has mean 0 and variance n, and we can approximate X as a Gaussian with fluctuations $\sqrt{n}$, as $\Pr[X \geq n\delta] \approx e^{-\frac{(n\delta)^2}{2n}} = e^{-\frac{n\delta^2}{2}}$. The bound we get from (12.216) is

$$\Pr[X \geq n\delta] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda n\delta}} \leq e^{\frac{n\lambda^2}{2} - \lambda n\delta} \tag{12.219}$$

[13] Independence can be relaxed to negative dependence, where if one increases, the rest have to stay the same or go down. There is also a martingale assumption. Independence is the baseline people have generalized, and for today we will use independence.

[14] These are called Rademacher random variables. A Bernoulli random variable takes values in $\{0,1\}$ instead.

[15] Compare the Taylor expansions term by term: $\cosh\lambda = \sum_{k\geq 0}\frac{\lambda^{2k}}{(2k)!}$ and $e^{\lambda^2/2} = \sum_{k\geq 0}\frac{\lambda^{2k}}{2^k k!}$, and $(2k)! = k!\,(k+1)(k+2)\cdots(2k) \geq 2^k k!$ since each of the k factors $k+1,\ldots,2k$ is at least 2.

for every $\lambda > 0$. Choosing $\lambda = 0$ or $\lambda = 100000$ (very big) results in lousy bounds: $\lambda = 0$ gives the trivial bound 1, and the bound blows up as $\lambda \to \infty$. The bound is tightest when $\lambda = \delta$, giving

$$\Pr[X \geq n\delta] \leq e^{-\frac{n\delta^2}{2}}, \tag{12.220}$$

which is a pretty sharp approximation (it matches the Gaussian approximation). This is called a Chernoff bound. For bounded variables like these it is also called a Hoeffding bound. Thus, sums of independent random variables are exponentially less likely to be far away from their mean.

We may ask what happens if we didn't assume a specific distribution on X_i, but instead just used the fact that they were all the same. That is, what information do we need about the X_i to get a bound like (12.220)? First, we want $\mathbb{E}[X_i] = 0$, so that $\mathbb{E}[e^{\lambda X_i}] = 1 + O(\lambda^2)$. A common assumption is that $|X_i| \leq 1$ is bounded, which (together with $\mathbb{E}[X_i] = 0$) means $\mathbb{E}[e^{\lambda X_i}] \leq e^{\lambda^2/2}$ (Hoeffding's lemma), so (12.220) still holds. The course web page links to other lecture notes (18.S997 High-dimensional statistics, Chapter 1) with more general conditions such as subgaussian and subexponential. Subgaussian says $\Pr[|X_i| \geq a] \leq c_1 e^{-c_2 a^2}$ for large enough a (Gaussian tails). Subexponential says $\Pr[|X_i| \geq a] \leq c_1 e^{-c_2 a}$ (exponential tails). Subgaussian is equivalent (for $\mathbb{E}[X_i] = 0$) to $\mathbb{E}[e^{\lambda X_i}] \leq e^{c\lambda^2}$ for all λ. Subexponential is equivalent to saying the moment generating function is finite for $|\lambda| < c$ (for some $c > 0$).

These are some of the standard assumptions people use. From weakest to strongest: the expectation is finite, then the variance is finite, then the third or fourth moment is finite, then all the moments are finite, then subexponential tails, then subgaussian tails.

12.2 Vector random variables

We want to model the uniform distribution over $|\psi\rangle \in S^d = S(\mathbb{C}^d)$, the unit sphere. The key property is invariance under unitary rotation $|\psi\rangle \mapsto U|\psi\rangle$: for every unitary U, $U|\psi\rangle$ has the same distribution as $|\psi\rangle$. On the sphere this property alone pins down the distribution, since there is only one rotation-invariant probability distribution on S^d. We write

$$\mathrm{Unif}(S^d) = \mathrm{Haar}(S^d), \tag{12.221}$$

which is also called the Haar measure over the sphere. As in Section 12.1, concentration bounds come from moments, so our goal is to compute the moments of $|\psi\rangle$, that is, the mixed moments of its coordinates $\psi_x = \langle x|\psi\rangle$. Since the coordinates are complex, these are expectations of products of coordinates and their conjugates. We call one with k coordinates and l conjugates a (k, l) moment. All the (n, n) moments are the entries of a single $d^n \times d^n$ matrix:

$$\mathbb{E}\left[|\psi\rangle\langle\psi|^{\otimes n}\right] = \sum_{x_1,\ldots,x_n,y_1,\ldots,y_n} \mathbb{E}\left[\psi_{x_1}\cdots\psi_{x_n}\psi^*_{y_1}\cdots\psi^*_{y_n}\right] |x_1,\ldots,x_n\rangle\langle y_1,\ldots,y_n|. \tag{12.222}$$

One mild difficulty is that the d coordinates of $|\psi\rangle$ are not independent (due to unit length), so these mixed moments don't factor over coordinates. (This is a negative dependence, as increasing one variable decreases the values of others due to the norm condition, so Chernoff bounds still work.) Since the coordinates on the sphere can never be independent, we instead work in all of $\mathbb{C}^d$.

So we want rotation symmetry and independent coordinates in $\mathbb{C}^d$? There are many distributions that are rotationally symmetric (like $\mathrm{Haar}(S^d)$), but these are not coordinate independent.

One example of both is deterministically giving the zero vector. It turns out the only other distributions satisfying these conditions are the complex Gaussians (this is the Herschel–Maxwell

theorem)

$$|g\rangle = \sum_{x=1}^{d} g_x |x\rangle, \qquad g_x = a_x + ib_x, \qquad a_x, b_x \sim N(0, \sigma^2) \text{ independently}, \tag{12.223}$$

for any variance σ^2. The coordinates are independent by construction; to see rotation symmetry, note that the density is

$$p(|g\rangle) = \left(2\pi\sigma^2\right)^{-d} e^{-\langle g|g\rangle/(2\sigma^2)}, \tag{12.224}$$

which is rotationally symmetric, since it depends only on $\langle g|g\rangle$. Every such Gaussian has $\mathbb{E}[|g\rangle] = 0$ and $\mathbb{E}[\langle g|g\rangle] = \sum_x \mathbb{E}[a_x^2 + b_x^2] = 2d\sigma^2$. We can choose the variance to be whatever we want, and we normalize $\mathbb{E}[\langle g|g\rangle] = 1$, the same as for a unit vector, by setting $\sigma^2 = \frac{1}{2d}$. Then the density becomes $p(|g\rangle) = \left(\frac{d}{\pi}\right)^d e^{-d\langle g|g\rangle}$. Let us calculate the $(1,1)$ moments $\mathbb{E}[g_x g_y^*]$, since the coordinates are independent with mean 0, the off-diagonal moments vanish, and diagonal terms have $\mathbb{E}[|g_x|^2] = \mathbb{E}[a_x^2 + b_x^2] = 2\sigma^2 = \frac{1}{d}$. Thus

$$\mathbb{E}[g_x g_y^*] = \delta_{xy}\, \mathbb{E}[|g_x|^2] = \frac{\delta_{xy}}{d}, \qquad \text{i.e.} \qquad \mathbb{E}[|g\rangle \langle g|] = \frac{I}{d}. \tag{12.225}$$

Because the $|g\rangle$ are rotationally invariant, the length and direction are independent, which gives a way of generating uniform vectors on the sphere:

$$|u\rangle = \frac{|g\rangle}{\sqrt{\langle g|g\rangle}} \sim \mathrm{Unif}(S^d). \tag{12.226}$$

This means to compute moments of uniformly random unit vectors, we can do a lot by computing things on Gaussians and then rescaling.

12.3 Higher moments

We want to compute higher moments, such as the $(2,1)$ moment $\mathbb{E}[g_{x_1} g_{x_2} g_{y_1}^*]$. Note that the (k,l) moments form the matrix $\mathbb{E}\left[|g\rangle^{\otimes k} \langle g|^{\otimes l}\right]$. It turns out this specific moment is always zero. The rough intuition is that for each (x_1, x_2, y_1), one of the x_i will be independent from y_1 and it must have mean 0, making the moment 0. The same argument shows that[16]

$$\mathbb{E}\left[|g\rangle^{\otimes k} \langle g|^{\otimes l}\right] = 0 \qquad \text{if } k \neq l. \tag{12.227}$$

Let us focus on $k = l$. For $k = 1$ we computed $\mathbb{E}[|g\rangle \langle g|] = \frac{I}{d}$, so let us move to $k = 2$:

$$\mathbb{E}[|g\rangle \langle g| \otimes |g\rangle \langle g|] = \sum_{x_1, x_2, y_1, y_2} \mathbb{E}[g_{x_1} g_{x_2} g_{y_1}^* g_{y_2}^*]\, |x_1, x_2\rangle \langle y_1, y_2|. \tag{12.228}$$

By the argument above, a moment can be nonzero only if no coordinate is left unpaired, so the phase of each g_{x_i} must cancel with the phase of some $g_{y_j}^*$ with $x_i = y_j$. For $k = 2$, x_1 can cancel with either y_1 or y_2 (and x_2 cancels with the other). In general, the nonzero (n,n) moments

[16] More rigorously, if $k \neq l$, then for each tuple $x_1, \ldots, x_k, y_1, \ldots, y_l$ there is some coordinate z that appears a times among the x_i and $b \neq a$ times among the y_j. Since the coordinates are independent, the moment factors as $\mathbb{E}\left[g_z^a (g_z^*)^b\right]$ times a moment of the other coordinates. This factor is zero, since g_z and $e^{i\phi} g_z$ have the same distribution, but the factor picks up a phase $e^{i(a-b)\phi}$.

correspond to permutations $\pi \in S_n$ with $x_i = y_{\pi(i)}$. This is Wick's theorem, and for $k = 2$ the two permutations are the identity and the swap, giving

$$\mathbb{E}[g_{x_1} g_{x_2} g^*_{y_1} g^*_{y_2}] = \mathbb{E}[g_{x_1} g^*_{y_1}]\,\mathbb{E}[g_{x_2} g^*_{y_2}] + \mathbb{E}[g_{x_1} g^*_{y_2}]\,\mathbb{E}[g_{x_2} g^*_{y_1}] \tag{12.229}$$

$$= \frac{\delta_{x_1 y_1}\delta_{x_2 y_2} + \delta_{x_1 y_2}\delta_{x_2 y_1}}{d^2}. \tag{12.230}$$

In terms of operators, this gives

$$\mathbb{E}[|g\rangle\langle g| \otimes |g\rangle\langle g|] = \frac{I + F}{d^2}, \tag{12.231}$$

where $F(|a\rangle \otimes |b\rangle) = |b\rangle \otimes |a\rangle$ is the swap operator (Figure 12.23). The same argument applies to higher moments, where

$$\mathbb{E}[g_{x_1} \cdots g_{x_n} g^*_{y_1} \cdots g^*_{y_n}] = \frac{1}{d^n} \sum_{\pi \in S_n} \prod_{i=1}^{n} \delta_{x_i, y_{\pi(i)}}, \qquad \mathbb{E}\big[|g\rangle\langle g|^{\otimes n}\big] = \frac{1}{d^n} \sum_{\pi \in S_n} P_\pi, \tag{12.232}$$

where P_π is the operator permuting n systems:

$$P_\pi = \sum_{x_1, \ldots, x_n} |x_1, \ldots, x_n\rangle\langle x_{\pi(1)}, \ldots, x_{\pi(n)}|. \tag{12.233}$$

Here $\pi \in S_n$ and $P_\pi \in L((\mathbb{C}^d)^{\otimes n})$, so $\pi \mapsto P_\pi$ is a representation of S_n.

Now back to $|u\rangle$. Let $|g\rangle = r\,|u\rangle$, where $r = \sqrt{\langle g|g\rangle}$ is independent of $|u\rangle$ by rotational invariance. Then,

$$\mathbb{E}\big[|g\rangle\langle g|^{\otimes n}\big] = \mathbb{E}[r^{2n}]\,\mathbb{E}\big[|u\rangle\langle u|^{\otimes n}\big], \tag{12.234}$$

where $\mathbb{E}[r^{2n}]$ is a scalar and $\mathbb{E}[|u\rangle\langle u|^{\otimes n}]$ has trace 1 since $|u\rangle$ is a unit vector. Thus,

$$\mathbb{E}\big[|u\rangle\langle u|^{\otimes n}\big] = \frac{\sum_\pi P_\pi}{\operatorname{tr} \sum_\pi P_\pi}. \tag{12.235}$$

For example, $n = 1$ gives $\mathbb{E}[|u\rangle\langle u|] = \frac{I}{d}$. For $n = 2$, we get $\frac{I+F}{d^2+d}$ (since $\operatorname{tr}(F) = \sum_{xy} \delta^2_{xy} = d$; see Figure 12.23).

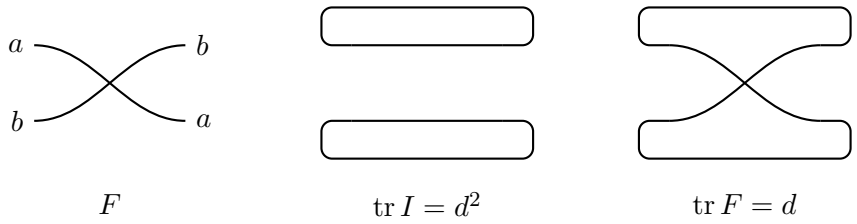


Figure 12.23: Left: the swap operator F drawn as two crossing wires, $F(|a\rangle \otimes |b\rangle) = |b\rangle \otimes |a\rangle$. Taking the trace connects each output wire back to its input. Each closed loop contributes a factor of d. For I (middle) we get two loops, so $\operatorname{tr} I = d^2$. For F (right), unlooping the wires gives a single circle, so $\operatorname{tr} F = d$.

To interpret this, look at the eigenspaces of F, which has eigenvalues ± 1. The $+1$ eigenspace is the symmetric subspace, spanned by the $\binom{d}{2}$ vectors $\frac{|x,y\rangle + |y,x\rangle}{\sqrt{2}}$ with $x < y$, together with the d vectors $|x, x\rangle$, for a total of $\frac{d(d+1)}{2}$. The -1 eigenspace is spanned by the $\binom{d}{2} = \frac{d(d-1)}{2}$ vectors

$\frac{|x,y\rangle - |y,x\rangle}{\sqrt{2}}$ with $x < y$. So $\Pi = \frac{1}{2}(I + F)$ is the projector onto the symmetric subspace, and we can write

$$\mathbb{E}\left[|u\rangle \langle u|^{\otimes 2}\right] = \frac{\frac{1}{2}(I+F)}{\frac{1}{2}d(d+1)} = \frac{\Pi}{\operatorname{tr}\Pi}, \tag{12.236}$$

which is the normalized projector onto the symmetric subspace.

For general n, we define

$$\operatorname{Sym}^n \mathbb{C}^d = \left((\mathbb{C}^d)^{\otimes n}\right)^{S_n} = \left\{|\psi\rangle \in (\mathbb{C}^d)^{\otimes n} : P_\pi |\psi\rangle = |\psi\rangle \ \ \forall \pi \in S_n\right\}. \tag{12.237}$$

Define $\Pi_{\text{sym}}^{(n,d)}$ to be the projector onto $\operatorname{Sym}^n \mathbb{C}^d$.

Lemma 12.3.1. $\Pi_{\text{sym}}^{(n,d)} = \dfrac{1}{n!} \displaystyle\sum_{\pi \in S_n} P_\pi$.

The proof was deferred in class. Here is a sketch.

Proof. Let $A = \frac{1}{n!} \sum_\pi P_\pi$. If $|\psi\rangle \in \operatorname{Sym}^n \mathbb{C}^d$, every P_π fixes it, so $A|\psi\rangle = |\psi\rangle$. For any $|\phi\rangle$ and $\sigma \in S_n$, $P_\sigma A |\phi\rangle = \frac{1}{n!} \sum_\pi P_{\sigma\pi} |\phi\rangle = A|\phi\rangle$, so everything A outputs is in $\operatorname{Sym}^n \mathbb{C}^d$. Hence $A^2 = A$ with image $\operatorname{Sym}^n \mathbb{C}^d$, and since $P_\pi^\dagger = P_{\pi^{-1}}$, A is Hermitian. So A is the orthogonal projector onto $\operatorname{Sym}^n \mathbb{C}^d$. □

Then

$$\mathbb{E}\left[|u\rangle \langle u|^{\otimes n}\right] = \frac{\Pi_{\text{sym}}}{\operatorname{tr}\Pi_{\text{sym}}} = \frac{\sum_{\pi \in S_n} P_\pi}{d(d+1)\cdots(d+n-1)}, \tag{12.238}$$

compared with the denominator d^n for Gaussians. The higher moments of $|u\rangle$ are lower than the Gaussian ones because Gaussians have a chance of being pretty big. This uses the fact that $\operatorname{tr}\Pi_{\text{sym}} = \binom{d+n-1}{n}$, so $\operatorname{tr}\sum_\pi P_\pi = n!\binom{d+n-1}{n} = d(d+1)\cdots(d+n-1)$. This is not the most surprising, since this relates to type classes. For each type t of strings in $[d]^n$, the vector $\sum_{x:\text{type}(x)=t} |x\rangle$ is symmetric. Different types give orthogonal vectors, and there are $\binom{d+n-1}{n}$ types. In fact these vectors form a basis of $\operatorname{Sym}^n \mathbb{C}^d$. We haven't proved that they span it, but this makes the dimension count plausible.

8.372 Quantum Information Science III **Fall 2026**

Lecture 13: October 27, 2026

Scribe: Aarush Vailaya *Random states, Rényi entropies, and entanglement*

13.1 Why random states and unitaries

Why are we studying random unitaries? Four justifications:

1. **Constructions.** Random unitaries are a source of useful constructions for quantum Shannon theory (like random codebooks for classical Shannon theory).
2. **Generic properties.** For any set of objects, we want to know which properties are generic and which are fine-tuned, so what are the generic properties of states? (For example, generic states are highly entangled.)
3. **Symmetry.** There is a connection to symmetry and representation theory.
4. **Pseudorandomness.** Even though these states are physically unrealistic, we will later see that there are realistic ensembles of states and unitaries that mimic many key properties. These are called pseudorandom states.

Recall that we proved last time

$$\mathbb{E}_{|u\rangle\in\mathbb{C}^d} |u\rangle\langle u|^{\otimes n} = \frac{\Pi_{\text{sym}}^{(n,d)}}{\operatorname{tr}\Pi_{\text{sym}}^{(n,d)}} = \frac{\frac{1}{n!}\sum_{\pi\in S_n} P_\pi}{\binom{d+n-1}{n}} = \frac{\sum_{\pi\in S_n} P_\pi}{d(d+1)\cdots(d+n-1)}. \tag{13.239}$$

For example, for $d = n = 2$ the symmetric subspace is spanned by $|00\rangle$, $|11\rangle$ and $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$, so

$$\mathbb{E}_u |u\rangle\langle u|^{\otimes 2} = \frac{1}{3}\Big(|00\rangle\langle 00| + |11\rangle\langle 11| + \frac{(|01\rangle + |10\rangle)(\langle 01| + \langle 10|)}{2}\Big). \tag{13.240}$$

13.2 Anticoncentration

Now for applications. The first is anticoncentration. Take a uniformly random unit vector $|u\rangle \in \mathbb{C}^d$ and measure it in the computational basis. Outcome z occurs with probability $p(z) = |\langle z|u\rangle|^2$. Then, $\mathbb{E}_u p(z) = \frac{1}{d}$ by symmetry. Alternatively,

$$\mathbb{E}_u p(z) = \operatorname{tr}\big[|z\rangle\langle z|\ \mathbb{E}_u |u\rangle\langle u|\big] = \operatorname{tr}\Big[|z\rangle\langle z| \cdot \frac{I}{d}\Big] = \frac{1}{d}. \tag{13.241}$$

Now, let us look at

$$\mathbb{E}_u p(z)^2 = \mathbb{E}_u |\langle z|u\rangle|^4 = \operatorname{tr}\big[|z\rangle\langle z| \otimes |z\rangle\langle z| \cdot \mathbb{E}_u |u\rangle\langle u| \otimes |u\rangle\langle u|\big]. \tag{13.242}$$

Two things happened here that are very useful tricks. The first was turning a power into the trace of a tensor power, which linearizes things: $|\langle z|u\rangle|^4$ is a degree-4 polynomial in the entries of $|u\rangle$

and $\langle u|$, but a linear function of $|u\rangle\langle u|^{\otimes 2}$. Second, this lets us pull apart the random part from the fixed part, which by (13.239) with $n = 2$ gives

$$\mathbb{E}_u p(z)^2 = \operatorname{tr}\left[|z\rangle\langle z| \otimes |z\rangle\langle z| \cdot \frac{I+F}{d(d+1)}\right] = \frac{2}{d(d+1)}. \tag{13.243}$$

One way to see this is that the above is $\operatorname{tr}\left(|z\rangle\langle z|^{\otimes 2} \frac{\Pi_{\text{sym}}}{\operatorname{tr}\Pi_{\text{sym}}}\right)$ and $|z\rangle\langle z|^{\otimes 2}$ lives in the symmetric subspace, so it equals $1/\operatorname{tr}\Pi_{\text{sym}} = \frac{2}{d(d+1)}$.

If $|u\rangle$ were a uniformly random computational basis state, $p(z)$ would be 1 with probability $\frac{1}{d}$ and 0 otherwise, so $\mathbb{E}p(z)^2$ would be $\frac{1}{d}$. But the actual value is much closer to $\frac{1}{d^2}$. If u were the uniform superposition, then this would be exactly $\frac{1}{d^2}$. This suggests that the above is some measure of how mixed the distribution p is. In particular, we denote

$$\sum_z p(z)^2 \tag{13.244}$$

as the collision probability.

13.3 Rényi entropies

The collision probability is related to Rényi entropy.

Definition 13.3.1 (Rényi entropy)**.**

$$H_\alpha(p) = \frac{1}{1-\alpha}\log\sum_x p(x)^\alpha = \frac{\alpha}{1-\alpha}\log\|p\|_\alpha, \qquad S_\alpha(\rho) = \frac{1}{1-\alpha}\log\operatorname{tr}\rho^\alpha. \tag{13.245}$$

The normalization factor at the front gives $H_\alpha(\text{unif}_d) = \log d$ for any α.

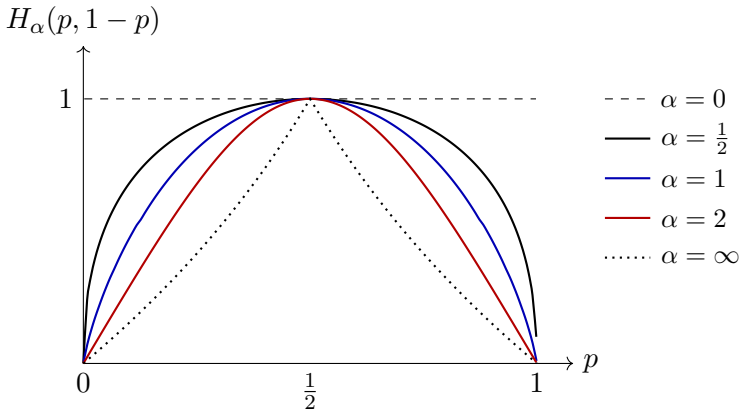


Figure 13.24: Rényi entropies of a bit, $H_\alpha(p, 1-p)$, for $\alpha = 0, \frac{1}{2}, 1, 2, \infty$. All are 0 for deterministic distributions and $\log 2 = 1$ for the uniform one, and they decrease as α increases. The $\alpha = \infty$ curve, $-\log\max(p, 1-p)$, is not concave.

A few basic properties are: $0 \le H_\alpha(p) \le \log d$, with 0 achieved for deterministic distributions (point masses) and $\log d$ for the uniform distribution; H_α is concave for $\alpha \le 1$; and it is a decreasing function of α (Figure 13.24).

For $S_0(\rho)$, taking the limit as $\alpha \to 0$ sends $p(x)^\alpha \to 1$ if $p(x) \neq 0$ and $p(x)^\alpha \to 0$ if $p(x) = 0$, which means $S_0(\rho) = \log \operatorname{rank} \rho$. Unlike von Neumann entropy, which has Fannes' inequality (saying if we perturb the state by a little, the entropy doesn't change a lot), here a small perturbation of the state can change $S_0(\rho)$ from 0 (a pure state) to $\log d$ (full rank). To fix this, sometimes people work with the smooth Rényi entropy.

Definition 13.3.2 (Smooth Rényi entropy)**.**

$$S_\alpha^\epsilon(\rho) = \begin{cases} \min\{S_\alpha(\sigma) : \frac{1}{2}\|\rho - \sigma\|_1 \leq \epsilon\} & \alpha < 1, \\ \max\{S_\alpha(\sigma) : \frac{1}{2}\|\rho - \sigma\|_1 \leq \epsilon\} & \alpha > 1. \end{cases} \tag{13.246}$$

Returning to Rényi entropy, if $\alpha > 0$, the function is continuous. For $\alpha = 1$, we need to take a limit, since plugging in $\alpha = 1$ gives $\frac{1}{0} \log \operatorname{tr} \rho = \frac{0}{0}$. By L'Hôpital's rule,

$$\lim_{\alpha \to 1} \frac{\log \operatorname{tr} \rho^\alpha}{1 - \alpha} = \lim_{\alpha \to 1} \frac{\operatorname{tr}(\rho^\alpha \log \rho)/\operatorname{tr} \rho^\alpha}{-1} = -\operatorname{tr} \rho \log \rho, \tag{13.247}$$

which gives $S_1(\rho) = S(\rho)$.

If $\alpha = 2$, we get $S_2(\rho) = -\log \operatorname{tr} \rho^2$, which for a distribution is $-\log$(collision probability). For quantum states, $\operatorname{tr} \rho^2$ is called the purity. Thus, we argued

$$\mathbb{E}_u \, 2^{-H_2(p)} = \mathbb{E}_u \sum_z p(z)^2 = d \cdot \frac{2}{d(d+1)} = \frac{2}{d+1}. \tag{13.248}$$

Professor Harrow gave a talk at Princeton hosted by a very good computer scientist who said they hated logs, and quantum information science had a lot of logs. If we don't want to take logs of a matrix, S_α with integer values of α is very nice, since we are just dealing with polynomials of our matrix elements. Two other useful cases are $S_\infty(\rho) = -\log \|\rho\|$,[17] and $S_{1/2}(\rho) = \log d + \log F(\rho, I/d)$, where $F(\rho, \sigma) = \left\|\sqrt{\rho}\sqrt{\sigma}\right\|_1^2$.

Starting from the maximally mixed state, to go from $\log d$ to $\frac{1}{2} \log d$ for von Neumann entropy, we have to move very far. However, for $S_\infty(\rho)$ to be $\frac{1}{2} \log d$, we only have to move to

$$\rho = \frac{1}{\sqrt{d}} |0\rangle\langle 0| + \frac{1 - 1/\sqrt{d}}{d - 1} \big(I - |0\rangle\langle 0|\big), \tag{13.249}$$

whose largest eigenvalue is $\frac{1}{\sqrt{d}}$ and whose other eigenvalues are slightly lower than $\frac{1}{d}$. Its trace distance from I/d is $\frac{1}{\sqrt{d}} - \frac{1}{d} \approx \frac{1}{\sqrt{d}}$, which is very small. While only $S_0(\rho)$ is infinitely unstable, all the Rényi entropies except $S_1(\rho)$ are unstable in that a small perturbation can create a big change.

Suppose $n_1 \gg n_2$ and

$$\rho = \pi |0\rangle\langle 0| \otimes \left(\frac{I}{2}\right)^{\otimes n_1} + (1 - \pi) |1\rangle\langle 1| \otimes \left(\frac{I}{2}\right)^{\otimes n_2} \otimes |0\rangle\langle 0|^{\otimes (n_1 - n_2)}. \tag{13.250}$$

The first term contributes 2^{n_1} eigenvalues equal to $\pi 2^{-n_1}$, and the second contributes 2^{n_2} eigenvalues equal to $(1 - \pi)2^{-n_2}$. Then

$$S_0(\rho) = \log(2^{n_1} + 2^{n_2}) \approx n_1, \qquad S_1(\rho) = H_2(\pi) + \pi n_1 + (1 - \pi) n_2, \tag{13.251}$$

[17] $\|\rho\|$ is the operator norm of ρ, the largest eigenvalue.

where $H_2(\pi)$ is the binary entropy (not the Rényi-2 entropy). Since $n_1 \gg n_2$, the largest eigenvalue is $(1-\pi)2^{-n_2}$, so

$$S_\infty(\rho) = n_2 - \log(1-\pi). \tag{13.252}$$

It turns out S_α for $\alpha < 1$ practically depends on n_1 only, and for $\alpha > 1$ practically depends on n_2 only. So, S_0 is sensitive to the small eigenvalues spread out over the big tail, while S_∞ is sensitive to the large eigenvalues with less entropy.

13.4 Entropy of random states

Going back to the problem, (13.248) says $\mathbb{E}_u 2^{-H_2(p)} = \frac{2}{d+1}$, meaning

$$\mathbb{E}H_2(p) = \mathbb{E}\Big[-\log\sum_z p(z)^2\Big] \geq -\log\mathbb{E}\sum_z p(z)^2 = \log\frac{d+1}{2} \geq \log d - 1 = n-1 \quad \text{if } d = 2^n, \tag{13.253}$$

using Jensen's inequality ($-\log$ is convex). Since H_α is decreasing in α, $H(p) \geq H_2(p)$, so $\mathbb{E}H(p) \geq n-1$, which means a random state measured in any fixed basis is nearly uniformly distributed.

This was used to argue for the hardness of random circuit sampling, where a quantum computer runs a random circuit of some depth and outputs samples from its output distribution p. This was an early candidate for a task that is easy for quantum computers but hard for classical ones. One part of the argument for why this is hard is anticoncentration: p is spread out over many strings, as the bound above shows. It is important that p is not uniform, since if it were, a classical computer could sample from it by flipping coins, and the quantum computer would not be doing anything interesting.

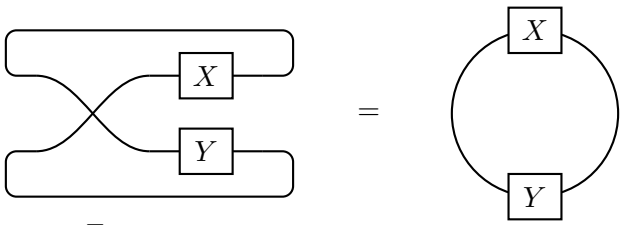
A similar argument shows most pure states are highly entangled. Suppose $|\psi\rangle \in \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$ is a random pure state of $n_A + n_B$ qubits, so $d_A = 2^{n_A}$ and $d_B = 2^{n_B}$. Then,

$$\mathbb{E}S(A) \geq \mathbb{E}S_2(A) = -\mathbb{E}\log\operatorname{tr}\psi_A^2 \geq -\log\mathbb{E}\operatorname{tr}\psi_A^2. \tag{13.254}$$

We now want to use the replica trick, which says

$$\operatorname{tr}[F(X\otimes Y)] = \operatorname{tr} XY. \tag{13.255}$$

A nice proof of this is with tensor diagrams (Figure 13.25).



F $\qquad$ $\operatorname{tr} XY$

$\operatorname{tr}[F(X\otimes Y)]$

Figure 13.25: The replica trick (13.255) in tensor diagrams. Left: $X \otimes Y$ followed by the swap F, with the trace gluing each output leg back to its input leg. Right: unravelling the wires gives a single loop through X and Y, which is $\operatorname{tr} XY$.

So now

$$\operatorname{tr}\psi_A^2 = \operatorname{tr}\big[(\psi_{A_1}\otimes\psi_{A_2})F\big] \tag{13.256}$$
$$= \operatorname{tr}\big[(\psi_{A_1B_1}\otimes\psi_{A_2B_2})\cdot(F_{A_1A_2}\otimes I_{B_1B_2})\big]. \tag{13.257}$$

The rough idea is that tensor products give all bilinear functions: to get a degree-k polynomial in ρ, take k tensor copies and then do something linear. We bring B back in because we only know one thing, the moments of random pure states. By (13.239) with $n=2$ in dimension d_Ad_B, $\mathbb{E}[\psi_{A_1B_1}\otimes\psi_{A_2B_2}]$ is $\frac{I+F}{d_Ad_B(d_Ad_B+1)}$, where F swaps A_1B_1 with A_2B_2. Swapping these composite systems is the same as swapping $A_1\leftrightarrow A_2$ and $B_1\leftrightarrow B_2$, so $F=F_{A_1A_2}F_{B_1B_2}$. So,

$$\mathbb{E}\operatorname{tr}\psi_A^2 = \operatorname{tr}\big[\mathbb{E}(\psi_{A_1B_1}\otimes\psi_{A_2B_2})\cdot(F_{A_1A_2}\otimes I_{B_1B_2})\big] \tag{13.258}$$
$$= \operatorname{tr}\Big[\frac{I_{A_1A_2B_1B_2}+F_{A_1A_2}F_{B_1B_2}}{d_Ad_B(d_Ad_B+1)}(F_{A_1A_2}\otimes I_{B_1B_2})\Big] \tag{13.259}$$
$$= \frac{\operatorname{tr}F_{A_1A_2}\cdot\operatorname{tr}I_{B_1B_2}+\operatorname{tr}I_{A_1A_2}\cdot\operatorname{tr}F_{B_1B_2}}{d_Ad_B(d_Ad_B+1)} \tag{13.260}$$
$$= \frac{d_Ad_B^2+d_A^2d_B}{d_Ad_B(d_Ad_B+1)} = \frac{d_A+d_B}{d_Ad_B+1} \tag{13.261}$$

(using $F_{A_1A_2}^2=I$, $\operatorname{tr}F=d$ and $\operatorname{tr}I=d^2$). Going back, assuming $d_A\le d_B$, we get

$$\mathbb{E}S(A) \ge -\log\frac{d_A+d_B}{d_Ad_B+1} \ge -\log\Big(\frac{1}{d_A}+\frac{1}{d_B}\Big) \ge \log(d_A)-1 = n_A-1, \tag{13.262}$$

where the middle step drops the $+1$ in the denominator, and the last uses $\frac{1}{d_B}\le\frac{1}{d_A}$. So Alice's state is nearly maximally mixed. Since ψ is pure, $S(A)$ is its entropy of entanglement (Lecture 11), so ψ is within one ebit of maximally entangled. The minus 1 is a bit pessimistic, as we could also write

$$\mathbb{E}S(A) \ge \log(d_A) - \log\Big(1+\frac{d_A}{d_B}\Big) \ge \log(d_A) - \frac{d_A}{(\ln 2)\,d_B}, \tag{13.263}$$

which is much better if $\frac{d_A}{d_B}\ll 1$. Another way of doing this is with trace distance.

Trace distance involves a weird absolute value, but if we were to take the distance in the Schatten 2-norm (Frobenius norm), that would be easier to deal with. We are analyzing

$$\mathbb{E}\|\psi_A - I/d_A\|_2^2 = \mathbb{E}\operatorname{tr}(\psi_A-I/d_A)^2 = \mathbb{E}\operatorname{tr}\psi_A^2 - \frac{1}{d_A} = \frac{d_A^2-1}{d_A(d_Ad_B+1)} \approx \frac{1}{d_B}. \tag{13.264}$$

To make sense of this, let us convert it to the trace norm.

In problem set 7 we show this when $d=2$, but in general $\|v\|_1\le\sqrt{d}\|v\|_2$, so

$$\mathbb{E}\|\psi_A - I/d_A\|_1 \le \sqrt{d_A}\,\mathbb{E}\|\psi_A-I/d_A\|_2 \le \sqrt{d_A}\sqrt{\mathbb{E}\|\psi_A-I/d_A\|_2^2} \approx \sqrt{\frac{d_A}{d_B}}, \tag{13.265}$$

meaning if d_A is much smaller than d_B, then ψ_A is very close to maximally mixed. With random matrix theory, we can get sharper answers to this question, but what is nice about these answers is that they are very general (anything whose first two moments match the uniform distribution will have a similar bound).

Someone asked: we already had an entropy bound, so how come this bound is bad when $d_A=d_B$? It turns out entropy $n-1$ does not mean being close to maximally mixed. If one qubit is pure and the rest are maximally mixed, $\rho=|0\rangle\langle 0|\otimes(I/2)^{\otimes(n-1)}$, then $S(\rho)=n-1$ but $\frac{1}{2}\|\rho-I/2^n\|_1=\frac{1}{2}$.

Next time, we will do random unitaries.

8.372 Quantum Information Science III **Fall 2026**

Lecture 14: October 29, 2026

Scribe: Aarush Vailaya *The symmetric subspace and Schur–Weyl duality*

14.1 Coherent information

We start by revisiting coherent information. Coherent information is $S(B) - S(E)$: you turn the channel into an isometry which turns Alice's input A' into B and E, the state is globally pure, and state A is the reference system (Figure 14.26). This can be thought of as how much information Bob gets minus how much is already shared with Eve.

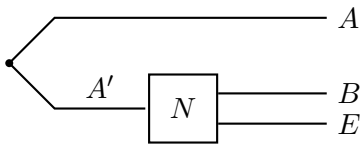


Figure 14.26: The coherent information $S(B) - S(E)$. Alice's input A' is purified by the reference system A, and the channel N is written as an isometry from A' to Bob's output B and Eve's environment E, so the state on ABE is pure.

Using this isometric channel for a perfect quantum channel ($N = \mathrm{id}$), Alice and Bob's state is the EPR pair while Eve's is nothing. If we have a classical channel,[18] putting in an EPR pair, we get out the GHZ state $\frac{1}{\sqrt{2}}(|000\rangle + |111\rangle)$. In both of these Bob's entropy is 1, but in the first example Eve's entropy is 0, while in the second example Eve's entropy is 1. Thus, the coherent information is 1 for the quantum channel and 0 for the classical channel. Thus, we can view coherent information (with many asterisks, like the many-copy limit and maximization over inputs) as the quantum channel capacity. It turns out the maximum will never be negative.

In the BB84 protocol, Alice sends qubits encoded in the X or Z basis, and Bob measures in the X or Z basis. There are a bunch of steps, but one of the steps is that Alice and Bob get a bunch of bits that are pretty well correlated, and Eve only knows a little bit of that. Perhaps they have n random bits, and Eve's share of that is $\frac{n}{4}$ bits. Then by using linear hashing (treating the bit string as a vector in $\mathbb{F}_2^n$ and applying a random matrix), they throw away information (going to a fewer number of bits) but wipe out Eve's knowledge. What happens is very similar to the coherent information formula, where they start off with $S(B)$ correlation, but must throw away some of it to decouple Eve.

[18] The isometry for a classical channel sends $|0\rangle \to |00\rangle$ and $|1\rangle \to |11\rangle$, copying the computational basis into B and E. Equivalently, we tensor with an environment E in the state $|0\rangle$, apply a CNOT from A' to E, and then discard E, which leaves the dephased state $\sum_x \langle x|\rho|x\rangle\ |x\rangle\langle x|$.

14.2 The symmetric subspace

First, we want to prove the following characterizations of the symmetric subspace:

$$\operatorname{Sym}^n \mathbb{C}^d = \operatorname{span}\{|\psi\rangle^{\otimes n}\} \tag{1}$$

$$= \operatorname{span}\Big\{|T_t\rangle = \binom{n}{t}^{-\frac{1}{2}} \sum_{\operatorname{type}(x)=t} |x\rangle : \sum_i t_i = n,\ t_1, \dots, t_d \in \mathbb{Z}_{\geq 0}\Big\} \tag{2}$$

$$= \big((\mathbb{C}^d)^{\otimes n}\big)^{S_n} = \big\{|\psi\rangle \in (\mathbb{C}^d)^{\otimes n} : P_d(\pi)\,|\psi\rangle = |\psi\rangle \ \ \forall \pi \in S_n\big\}. \tag{3}$$

We claim these all span the same subspace. The identity with (2) claims $\dim \operatorname{Sym}^n \mathbb{C}^d = \#\text{types} = \binom{n+d-1}{n}$.

First, we show (3) $\subseteq$ (2), meaning any symmetric vector can be written as a linear combination of $|T_t\rangle$. We use the fact that

$$\Pi_{\text{sym}} = \frac{1}{n!} \sum_{\pi \in S_n} P_\pi \tag{14.266}$$

projects onto $\operatorname{Sym}^n \mathbb{C}^d$, which has not been proven yet but will be soon (Proposition 14.3.1). Note (3) $= \operatorname{Im} \Pi_{\text{sym}}$ by (14.266), so for $|\psi\rangle \in$ (3), written in the computational basis as $|\psi\rangle = \sum_x c_x |x\rangle$, we have

$$|\psi\rangle = \Pi_{\text{sym}} |\psi\rangle = \sum_x c_x \underbrace{\Pi_{\text{sym}} |x\rangle}_{\alpha |T_t\rangle} \in (2). \tag{14.267}$$

To see why $\Pi_{\text{sym}} |x\rangle \propto |T_t\rangle$ with $t = \operatorname{type}(x)$, take for example $x = 3110012$. Every $P_\pi |x\rangle$ is a rearrangement of x, so it has the same type, and averaging over all 7! permutations hits each of the $\frac{7!}{2!\,3!\,1!\,1!} = 420$ rearrangements equally often ($2!\,3!\,1!\,1! = 12$ times each). So

$$\Pi_{\text{sym}} |3110012\rangle = \frac{12}{7!} \sum_{\operatorname{type}(y)=\operatorname{type}(3110012)} |y\rangle = \frac{1}{\sqrt{420}} |T_t\rangle . \tag{14.268}$$

In general $\alpha = \binom{n}{t}^{-1/2}$.

Next, we show (1) $\subseteq$ (3), which is easy since $P_\pi |\psi\rangle^{\otimes n} = |\psi\rangle^{\otimes n} \in$ (3), which linearly extends to all $|\Phi\rangle = \sum_i c_i |\Psi_i\rangle^{\otimes n} \in$ (1).

The hardest one is (2) $\subseteq$ (1), which requires showing every $|T_t\rangle$ is a linear combination of spin-coherent states (states of the form $|\psi\rangle^{\otimes n}$). Consider the state

$$\prod_{i=1}^d \left(\frac{d}{dx_i}\right)^{t_i} \overbrace{(x_1 |1\rangle + \cdots + x_d |d\rangle)^{\otimes n}}^{|\psi\rangle^{\otimes n} \in (1)} \propto |T_t\rangle . \tag{14.269}$$

Expanding, $(x_1 |1\rangle + \cdots + x_d |d\rangle)^{\otimes n} = \sum_y x_1^{s_1} \cdots x_d^{s_d} |y\rangle$ with $s = \operatorname{type}(y)$. Since $\sum_i s_i = \sum_i t_i = n$, the derivative kills every monomial except $s = t$, which becomes the constant $t_1! \cdots t_d!$. For example, consider $t = (2, 0, 1)$, so $n = d = 3$. Then $\frac{d^2}{dx_1^2} \frac{d}{dx_3}$ kills everything that is not proportional to $x_1^2 x_3$, and so we get

$$\begin{aligned} \frac{d^2}{dx_1^2} \frac{d}{dx_3} \big(x_1 |1\rangle + x_2 |2\rangle + x_3 |3\rangle\big)^{\otimes 3} &= \frac{d^2}{dx_1^2} \frac{d}{dx_3} \Big[x_1^2 x_3 \big(|113\rangle + |131\rangle + |311\rangle\big) + \cdots\Big] \\ &= 2\big(|113\rangle + |131\rangle + |311\rangle\big) = 2\sqrt{3}\ |T_{(2,0,1)}\rangle . \end{aligned} \tag{14.270}$$

The key point about derivatives is that taking derivatives of things in (1) gives a limit of linear combinations of elements in (1):

$$\frac{d}{ds}|\phi(s)\rangle = \lim_{\epsilon\to 0}\frac{|\phi(s+\epsilon)\rangle - |\phi(s)\rangle}{\epsilon}, \tag{14.271}$$

where each term on the right is in (1) for each ϵ. Since (1) is a finite-dimensional subspace, it is closed, so the limit is also in (1). So $|T_t\rangle \in (1)$.

Another reason (1) is the symmetric subspace is that $\int d\psi\ |\psi\rangle\langle\psi|^{\otimes n} \propto \Pi_{\text{sym}}$. We showed this in Lecture 12 without using the characterizations above. The image of this integral lies in $\text{span}\{|\psi\rangle^{\otimes n}\}$, while the image of Π_{sym} is all of (3), so $(3) \subseteq (1)$.

14.3 Representation theory

Another perspective is representation theory.

Definition 14.3.1 (Representation)**.** *A group representation (R, V) is a homomorphism $R : G \to U(V)$ (meaning $R(gh) = R(g)R(h)$).*

In general people consider $R : G \to \text{GL}_n$, but for any finite or compact group, we can get away with restricting to the unitaries. Two representations are equivalent, $(R_1, V_1) \cong (R_2, V_2)$, if there exists an invertible $T : V_1 \to V_2$ such that for all g, $R_2(g) = TR_1(g)T^{-1}$.

A representation (R, V) is reducible if it is equivalent to $(R_1 \oplus R_2, V_1 \oplus V_2)$ with $\dim V_1, \dim V_2 > 0$, i.e. for some T and all g

$$TR(g)T^{-1} = \begin{pmatrix} R_1(g) & 0 \\ 0 & R_2(g) \end{pmatrix}. \tag{14.272}$$

An irrep (or irreducible representation) is a representation that is not reducible. Irreps are really powerful, as each group has a set $\hat{G}$ of inequivalent irreducible representations (for a finite group there are finitely many, one for each conjugacy class), and every finite-dimensional representation (R, V) decomposes into the irreps. In particular,

$$V \cong \bigoplus_{\lambda\in\hat{G}} V_\lambda \otimes \mathbb{C}^{m_\lambda}, \tag{14.273}$$

which is akin to prime factorization: (r_λ, V_λ) is the irrep labeled by λ and $m_\lambda \geq 0$ is its multiplicity. Concretely, there must exist T such that

$$TR(g)T^{-1} = \bigoplus_\lambda r_\lambda(g) \otimes I_{m_\lambda} \quad \forall g \in G. \tag{14.274}$$

In physics, for rotations the irrep label λ is the total angular momentum, the rotations act on each irrep through r_λ, and the other degrees of freedom live in the multiplicity space $\mathbb{C}^{m_\lambda}$.

Going back to symmetric subspaces, $U \mapsto U^{\otimes n}$ is a representation of $U(d)$. ($U \mapsto U^*$ and $U \mapsto \det(U)$ are also representations, while $U \mapsto U^T$ and $U \mapsto U^\dagger$ are not, as they reverse the order of multiplication.) We are always talking about complex unitary finite-dimensional representations, and $U \mapsto U$ is called the defining representation.

Now, we want to decompose the representation $U \mapsto U^{\otimes n}$ into irreducible representations. One thing that becomes important is Schur–Weyl duality, which starts from the fact that $U \mapsto U^{\otimes n}$ and $\pi \mapsto P_d(\pi)$, which both act on $(\mathbb{C}^d)^{\otimes n}$ (n qudits), commute.

In particular, $U \mapsto U^{\otimes n}$ is reducible if there is a nonzero R-invariant subspace that is strictly smaller (since then we can block diagonalize; for a unitary representation, the orthogonal complement of an invariant subspace is also invariant). One good idea is to try the symmetric subspace. In particular, it turns out $U^{\otimes n} \operatorname{Sym}^n \mathbb{C}^d = \operatorname{Sym}^n \mathbb{C}^d$, which comes out of (1), since $U^{\otimes n} |\psi\rangle^{\otimes n} = (U|\psi\rangle)^{\otimes n}$. This means $(U^{\otimes n}, (\mathbb{C}^d)^{\otimes n})$ is reducible, but it turns out $(U^{\otimes n}, \operatorname{Sym}^n \mathbb{C}^d)$ is irreducible (we skip the proof for now).

Another irrep is $(U^{\otimes n}, \operatorname{Anti}^n \mathbb{C}^d)$ on the antisymmetric subspace

$$\operatorname{Anti}^n \mathbb{C}^d = \{|\psi\rangle : P_\pi |\psi\rangle = \operatorname{sign}(\pi) |\psi\rangle \ \ \forall \pi \in S_n\}, \tag{14.275}$$

where $\operatorname{sign}(\pi) = (-1)^{\#\text{transpositions needed to form } \pi}$. The dimension of this subspace is $\binom{d}{n}$. For $n = 2$, $(\mathbb{C}^d)^{\otimes 2} = \operatorname{Sym}^2 \mathbb{C}^d \oplus \operatorname{Anti}^2 \mathbb{C}^d$, but this is not true for $n > 2$. These kinds of decompositions are examples of Schur–Weyl duality. One more comment is that the equivalence is with respect to the action of $U(d)$, but the permutation group also does something here: the swap acts as $+1$ on $\operatorname{Sym}^2 \mathbb{C}^d$ and as -1 on $\operatorname{Anti}^2 \mathbb{C}^d$. The irreps of S_2 are the trivial irrep and the sign irrep (sending the swap to -1), which correspond to the two different spaces. It turns out the irreps of S_n describe how to decompose $U(d)$.

The two groups are closely related, so people like Richard Stanley in his *Enumerative Combinatorics* start with permutations, prove things about irreps using combinatorics, and then move to unitaries. Meanwhile, others starting from physics or Lie algebras start from unitaries, and S_n falls out as a corollary.

Now we go back to the fact (14.266) we used earlier. Let (R, V) be a representation and let

$$V^G = \{|\psi\rangle \in V : R(g) |\psi\rangle = |\psi\rangle \ \ \forall g \in G\}. \tag{14.276}$$

Proposition 14.3.1.

$$\Pi = \frac{1}{|G|} \sum_{g \in G} R(g) \tag{14.277}$$

is the orthogonal projector onto V^G. For compact groups the same holds for $\Pi = \int dg\, R(g)$, with dg the Haar measure.

This specializes to the claim on permutations in (14.266) with $G = S_n$ and $R = P_d$.

Proof. First, a lemma is $R(h)\Pi = \Pi$, since

$$R(h) \frac{1}{|G|} \sum_{g \in G} R(g) = \frac{1}{|G|} \sum_{g \in G} R(hg) = \frac{1}{|G|} \sum_{g \in hG} R(g) = \frac{1}{|G|} \sum_{g \in G} R(g) = \Pi. \tag{14.278}$$

To show Π is the projector we look at $\Pi^\dagger \Pi$:

$$\Pi^\dagger \Pi = \frac{1}{|G|} \sum_h R(h)^\dagger \Pi = \frac{1}{|G|} \sum_h R(h)^{-1} \Pi = \frac{1}{|G|} \sum_h R(h^{-1}) \Pi = \frac{1}{|G|} \sum_h \Pi = \Pi. \tag{14.279}$$

Here we use the fact that when talking about a representation, we mean a unitary representation, so $R(h)^\dagger = R(h)^{-1}$. The claim without proof is that for finite and compact groups, all non-unitary representations are equivalent to a unitary one.

Then, $\Pi^\dagger \Pi = \Pi$ means Π is a projector (taking $\dagger$ of both sides gives $\Pi^\dagger = \Pi$, and then $\Pi^2 = \Pi$). We now show it projects onto V^G. First, $\Pi |\psi\rangle$ lives in V^G (the invariant subspace) since $R(h)\Pi |\psi\rangle = \Pi |\psi\rangle$ by (14.278). Second, if $|\psi\rangle \in V^G$, we have

$$\Pi |\psi\rangle = \frac{1}{|G|} \sum_g R(g) |\psi\rangle = \frac{1}{|G|} \sum_g |\psi\rangle = |\psi\rangle. \tag{14.280}$$

□

Note that Π mapping everything into V^G and Π leaving everything in V^G alone is not enough to show Π is the orthogonal projector, since that only gives $\Pi^2 = \Pi$. For example, the (non-orthogonal) projector $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ squares to itself, maps everything into $\mathrm{span}\{|0\rangle\}$ and fixes $|0\rangle$, but it is not the orthogonal projector onto $|0\rangle$.

14.4 Schur–Weyl duality

So what is the point? We might believe that $(1), (2), (3)$ are equivalent, so here are a few more surprising consequences.

The proof in Section 14.2 works for any vector space, such as $M_d = \mathbb{C}^{d\times d}$. Here $\mathrm{Sym}^n M_d$ consists of the $Y \in M_d^{\otimes n}$ with $P_\pi Y P_\pi^\dagger = Y$ for all π, i.e. the operators commuting with every P_π. Then

$$\mathrm{Sym}^n M_d = \mathrm{span}\{X^{\otimes n} : X \in M_d\} = \mathrm{span}\{U^{\otimes n} : U \in U(d)\}, \tag{14.281}$$

where the last step has not been justified yet but does not require that much work. With a bit more work, this gives us the double-commutant theorem. Before stating it, let us go over some notation.

If $\mathcal{A} \subseteq M_d^{\otimes n}$ is a vector space closed under multiplication and under $\dagger$, and containing I, then it is called an $(*\text{-})$"algebra." The commutant is $\mathrm{Comm}(\mathcal{A}) = \{B : [A, B] = 0\ \forall A \in \mathcal{A}\}$.

Theorem 14.4.1 (Double-commutant theorem)**.** *For finite dimensions,* $\mathrm{Comm}(\mathrm{Comm}(\mathcal{A})) = \mathcal{A}$.

In particular, let $\mathcal{A} = \mathrm{span}\{P_d(\pi) : \pi \in S_n\}$ and $\mathcal{B} = \mathrm{span}\{U^{\otimes n} : U \in U(d)\}$. Then $\mathrm{Comm}(\mathcal{A}) = \mathcal{B}$ and $\mathrm{Comm}(\mathcal{B}) = \mathcal{A}$. The easy part is that the two commute, but the complicated part is that nothing else commutes with these. This comes from the symmetric subspace $\mathrm{Sym}^n M_d$, which is the set of things that commute with all the permutations; proving it is spanned by the unitaries (14.281) shows $\mathrm{Comm}(\mathcal{A}) = \mathcal{B}$. The double-commutant theorem lets us go in the other direction.

Let $(q_\lambda^d, Q_\lambda^d)$ be irreps of $U(d)$ and (p_λ, P_λ) be irreps of S_n, where λ ranges over Young diagrams with n boxes and at most d rows.

Theorem 14.4.2 (Schur–Weyl duality)**.**

$$(\mathbb{C}^d)^{\otimes n} \cong \bigoplus_\lambda Q_\lambda^d \otimes P_\lambda. \tag{14.282}$$

Just from the fact that $[U^{\otimes n}, P_d(\pi)] = 0$, we get that

$$U(d) \times S_n \to U((\mathbb{C}^d)^{\otimes n}), \quad (U, \pi) \mapsto U^{\otimes n} P_d(\pi) \tag{14.283}$$

is a representation (since the actions commute). Moreover, the irreps of $U(d) \times S_n$ are the tensor products of the irreps of each group. Thus, we get

$$(\mathbb{C}^d)^{\otimes n} \cong \bigoplus_{\lambda_1, \lambda_2} Q_{\lambda_1}^d \otimes P_{\lambda_2} \otimes \mathbb{C}^{M_{\lambda_1, \lambda_2}}. \tag{14.284}$$

The idea is that operators on the first tensor factor and operators on the second commute with each other, but they may also commute with operators on the third factor. But the double-commutant theorem is saying there are no more degrees of freedom left. This provides intuition for the double-commutant theorem: if $\mathcal{A}$ is all operations on the first 4 qubits of 10 qubits, then the commutant of $\mathcal{A}$ is all operations on the qubits that $\mathcal{A}$ does not touch (the last 6 qubits), and

the commutant of that returns back to $\mathcal{A}$. The key is that by just knowing they commute we have (14.284), but there might be additional degrees of freedom (the $\mathbb{C}^{M_{\lambda_1,\lambda_2}}$). In practice it turns out there is additional structure and this additional thing does not exist: $M_{\lambda_1,\lambda_2} = \delta_{\lambda_1,\lambda_2}$, which gives (14.282).