8.372 Quantum Information Science III (2024)

Lecturer: Aram W. Harrow
TA: Yuanjie (Collin) Ren
Scribes: Students of 8.372 Fall 2024

October 26, 2024

Contents

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.

8.372 Quantum Information Science III **Fall 2024**

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:

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}. \tag{2.13}$$

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[3], 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:

$$\|x\|_p = \max_{\|y\|_q = 1} |\langle x, y \rangle| . \tag{2.14}$$

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

$$\|x\|_2 = \max_{\|y\|_2 = 1} |\langle x, y \rangle| . \tag{2.15}$$

[3]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:

$$\begin{aligned} \|x\|_1 &= \max_{\|y\|_\infty = 1} |\langle x, y\rangle|, \\ \sum_i |x_i| &= \max_{\max(|y_i|)=1} |\langle x, y\rangle|. \end{aligned} \tag{2.16}$$

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:

$$\begin{aligned} \|x\|_\infty &= \max_{\|y\|_1 = 1} |\langle x, y\rangle|, \\ \max(x_i) &= \max_{\sum_i |y_i| = 1} |\langle x, y\rangle|. \end{aligned} \tag{2.17}$$

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:

$$\|X\|_{S_p} \equiv \|X\|_p = \|\Sigma(X)\|_p, \tag{2.18}$$

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:

$$\|X\|_\infty = \max \Sigma(X) = \max_{\|v\|_2 = 1} \|Xv\|. \tag{2.19}$$

S_1 and S_2 can be expressed without using the SVD:

$$\|X\|_1 = \mathrm{Tr}\sqrt{X^\dagger X}, \|X\|_2 = \sqrt{\mathrm{Tr}\, X^\dagger X}, \tag{2.20}$$

where the square root is well-defined because $X^\dagger X$ is positive semi-definite. It is easy to show that Eq. 2.20 is consistent with Eq. 2.18 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:

$$S = \{x : \|x\| = 1\}, \tag{2.21}$$

$$B = \{x : \|x\| \leq 1\}. \tag{2.22}$$

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:

$$T(p,q) = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\sum_i |p_i - q_i|\,. \tag{2.23}$$

Note that

$$T(p,q) = \frac{1}{2}\sum_i |p_i - q_i| \le \frac{1}{2}\sum_i (p_i + q_i) = 1, \tag{2.24}$$

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 $X = i$:

$$P(X \text{ from } p | X = i) = \frac{p_i}{p_i + q_i}. \tag{2.25}$$

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

$$P(\text{correct}|X=i) = \max\left(\frac{p_i}{p_i+q_i}, \frac{q_i}{p_i+q_i}\right) = \frac{1}{2} + \frac{|p_i - q_i|}{2(p_i+q_i)}. \tag{2.26}$$

Then the probability that we guess correctly in general is:

$$\begin{aligned} P(\text{correct}) &= \sum_i P(\text{correct}|X=i)P(X=i) \\ &= \sum_i \left[\frac{1}{2} + \frac{|p_i - q_i|}{2(p_i+q_i)}\right]\left[\frac{p_i+q_i}{2}\right] \\ &\frac{1}{2} + \frac{T(p,q)}{2}. \end{aligned} \tag{2.27}$$

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:

$$F(p,q) = \langle \sqrt{p}, \sqrt{q} \rangle \,, \tag{2.28}$$

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:

$$\begin{aligned} F(p_i, q_i) &= \sum_i \sqrt{p_i q_i} \\ &= \sum_{a,b} \sqrt{p_a p_b q_a q_b} \\ &= F(p_a, q_a) F(p_b, q_b). \end{aligned} \tag{2.29}$$

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

$$1 - F \leq T \leq \sqrt{2(1-F)}. \tag{2.30}$$

Even through T doesn't factorize, Eq. 2.30 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,*

$$T(\rho,\sigma) = \frac{1}{2} \|\rho - \sigma\|_1 \tag{2.31}$$

$$= \max_{0 \preceq M \preceq \mathbb{I}} \mathrm{tr}[M(\rho - \sigma)] \tag{2.32}$$

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 (2.32) 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

$$\max_{U,V} |\langle\phi^\rho| (\mathbb{I}\otimes U_B)(\mathbb{I}\otimes V_B) |\phi^\sigma\rangle| \tag{2.33}$$

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,

$$\begin{aligned}
(2.33) &= \max_U |\langle\phi^\rho| (\mathbb{I}\otimes U) |\phi^\sigma\rangle| && (2.34)\\
&= \max_U \langle\Gamma| (\sqrt{\rho}\otimes\mathbb{I})(\mathbb{I}\otimes U)(\sqrt{\sigma}\otimes\mathbb{I}) |\Gamma\rangle && (2.35)\\
&= \max_U \langle\Gamma| (\sqrt{\rho}\sqrt{\sigma})\otimes U |\Gamma\rangle && (2.36)\\
&= \max_U \mathrm{tr}\left[\sqrt{\rho}\sqrt{\sigma}U^\top\right] && (2.37)\\
&= \|\sqrt{\rho}\sqrt{\sigma}\|_1 && (2.38)
\end{aligned}$$

where in (2.37) 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*

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{3.39}$$

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 $0 \leq \lambda \leq 1$,

$$\lambda H(p_1) + (1-\lambda) H(p_2) \leq H[\lambda p_1 + (1-\lambda) p_2]. \tag{3.40}$$

3. *Norm power series expansion.*

$$\|p\|_{1+\epsilon} = \left(\sum_{x \in X} p(x)^{1+\epsilon} \right)^{\frac{1}{1+\epsilon}} = 1 + \epsilon H(p) + \mathcal{O}\big(\epsilon^2\big). \tag{3.41}$$

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*

$$H(X, Y) = -\sum_{(x,y)\in(X,Y)} p(x, y) \log_2 p(x, y). \tag{3.42}$$

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*

$$\begin{aligned} H(Y|X) &= -\sum_{x\in X} \mathbb{P}(X = x) H(Y|X = x) && (3.43)\\ &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log p(y|x) && (3.44)\\ &= H(X, Y) - H(X), && (3.45) \end{aligned}$$

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),*

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2). \tag{3.46}$$

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*

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{3.47}$$

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,*

$$I(X; Y) \geq 0 \iff H(X) + H(Y) \geq H(X, Y) \iff H(Y|X) \leq H(Y). \tag{3.48}$$

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*

$$H(\rho) = -\sum_k \lambda_k \log_2 \lambda_k \tag{3.49}$$

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,*

$$H(A) + H(ABC) \leq H(AB) + H(AC). \tag{3.50}$$

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*

$$C = \max_X I(X; Y), \tag{3.51}$$

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.

□

8.372 Quantum Information Science III **Fall 2024**

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

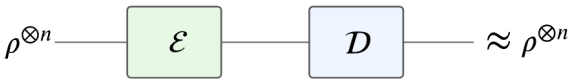
The projector P_{typ} has the following important properties:

- *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 $2^{nS(\rho)}$.

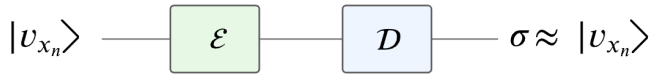
4.5 Quantum Compression

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.

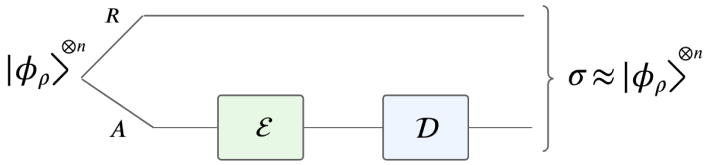


2. In the second scheme we can define x_n λ_n and check $\mathop{\mathbb{E}}_{x^n \sim \lambda^n} F(\sigma, |v_{x_n}\rangle) \approx 1$.



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.

8.372 Quantum Information Science III **Fall 2024**

Lecture 5: September 19, 2024

Scribe: John Blue and Jonathan Lu *Relative entropy*

5.1 Huffman codes as an interpretation of entropy

Here's an interesting interpretation of entropy given by Shannon. Suppose $X \sim p$. How surprised would you be to see that $X = x$? For example, in the English language, you wouldn't be very surprised if $X = \texttt{e}$, but you would be pretty surprised if $X = \texttt{q}$. Define

$$\text{surprise}(x) = \log \frac{1}{p(x)}. \tag{5.52}$$

The idea of this definition is that $1/p(x)$ gets larger as $p(x)$ gets smaller, so that as things are less probable we are more surprised. But why the log? This comes from an exp licit construction of an information compression scheme known as *Huffman* coding. The idea is to map an outcome x into a bitstring

$$x \longrightarrow \text{Enc}(x) \text{ s.t. } |\text{Enc}(x)| = \left\lceil \log \frac{1}{p(x)} \right\rceil = \lceil \text{surprise}(x) \rceil. \tag{5.53}$$

If you pretend for a moment that all the probabilities are dyadic (i.e. 2^{-k} for some k depending on x), then the log gives an immediate interpretation of representing a number as a bitstring. In making the encoding, you have to be careful to ensure that it can be decodable. One straightforward way to do this is by ensuring that the code is *prefix-free*, i.e. that no codeword is the prefix of another codeword. If this weren't the case, we would get lost trying to decode locally. As an example, consider the Table 5.1 below. If we instead encoded `a` with `1`, and we have a stream of

x	$p(x)$	$\text{Enc}(x)$
`a`	1/2	0
`b`	1/4	10
`c`	1/8	110
`d`	1/8	111

Table 5.1: Huffman coding for a dyadic distribution over 4 characters.

bits coming in that look like `111`, we couldn't distinguish `aaa` from `d`. (We could add separation characters between each encoded bitstring, but that would increase the encoding size!)

Note that if p is a dyadic distribution, $H(p) = \mathbb{E}[|\text{Enc}(X|]$, giving a constructive interpretation of entropy as the average Huffman encoding length. In general, the ceiling gives a few off-by-one errors that makes the Huffman code slightly more annoying to deal with. We won't get into that here, but it doesn't make any practical impact on the fundamental concepts we have discussed.

5.2 Relative entropy

Let's now imagine that we are trying to follow the Huffman procedure to encode our data into bits. The data comes from a distribution p, but we don't know p. Instead, we guess a distribution q and

encode according to q instead. How good is our Huffman code now? Define the Huffman encoding map using q as Enc_q. The new average length of the encoding is given by

$$\mathbb{E}_{X\sim p}[\text{Enc}_q(X)] = \sum_x p(x) \log \frac{1}{q(x)}. \tag{5.54}$$

To study this quantity, we would like to write it in terms of the actual entropy $H(p)$ and some kind of measure of how much q deviates from p.

Definition 5.2.1. *The relative entropy of q relative to p is given by $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.*

In a very loose sense, the relative entropy $D(p||q)$ is meant to give a "distance" between distributions p and q. However, note that $D(p||q)$ is *not* symmetric. Also, p is supported on a character in which q is not, $D(p||q)$ is infinite! So what can we say about it?

Theorem 5.2.1. $D(p||q) \geq 0$.

Proof. One way to prove this is by applying Shannon's noiseless coding theorem. But we'll do this by a direct algebraic proof because of how important this bound is. First, note that $1 + z \leq e^z \ \forall z \in \mathbb{R}$. In particular, $z \leq e^z - 1$. Now let $z = \ln y$, so that $\ln y \leq y - 1$ and thus $\ln \frac{1}{y} \geq 1 - y$. Hence,

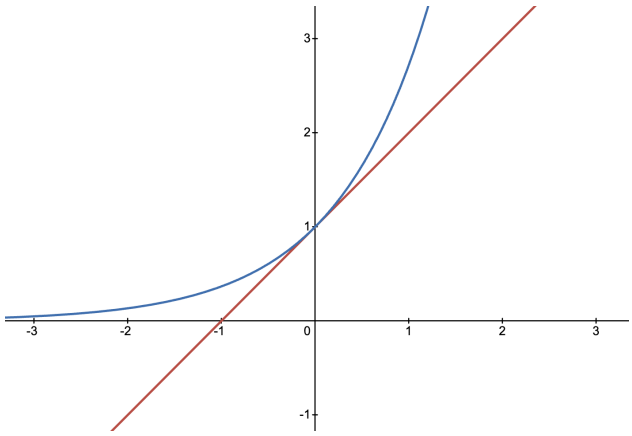


Figure 5.1: e^z (blue) is lower bounded by $1 + z$ (red).

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \frac{1}{\ln 2} \sum_x p(x) \ln \frac{p(x)}{q(x)} \tag{5.55}$$

$$\geq \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) = \sum_x p(x) - q(x) = 1 - 1 = 0. \tag{5.56}$$

□

There are a number of important corollaries that follow almost immediately from this result.

Corollary 5.2.1. *$H(p) \leq \log d$ where d is the size of the character set from which p draws.*

Proof. First, define $u \in \mathbb{R}^d$ to be the uniform distribution, that is $u = (1/d, \ldots, 1/d)$. Then

$$0 \leq D(p||u) = \sum_x p(x)(\log p(x) + \log d) \tag{5.57}$$

$$= \log d - H(p) \tag{5.58}$$

□

This quick proof emphasizes the the asymmetry of the relative entropy is telling you something - the mixed thing should always go second!

We give the following inequality without proof

Theorem 5.2.2 (Pinsker's Inequality)**.**

$$D(p||q) \geq \frac{1}{2\ln 2}||p-q||_1^2 \tag{5.59}$$

Using this, we can make rigorous an intuition that $H(p)$ being close to $\log d$ means that p is close to uniform. Suppose $H(p) \geq \log d - \delta$. Then $D(p||u) \leq \delta$, which by Pinsker's inequality implies that $||p-u||_1 \leq \sqrt{2\ln(2)\delta}$.

We'll next use relative entropy to prove things about mutual information.

Corollary 5.2.2.

$$0 \leq I(X;Y) = H(X) + H(Y) - H(XY) \tag{5.60}$$
$$= H(X) - H(X|Y) \tag{5.61}$$
$$= H(Y) - H(Y|X) \tag{5.62}$$

We can interpret this as saying that a) mutual information is a correlation (non-negative) b) conditioning reduces entropy.

Proof. Consider the relative entropy between a joint distribution p_{XY}, and the product of it's marginals:

$$D(p_{XY}||p_X \otimes p_Y) = \sum_{x,y} p_{XY}(x,y)\left(\log(p_{XY}(x,y)) - \log p_X(x) - \log p_Y(y)\right) \tag{5.63}$$

The first term is simply $-H(XY)$. Looking at the second term, $\sum_{x,y} p_{XY}(x,y)\log p_X(x)$, we see that by summing over y, we recover the marginal $p_X(x)$, and so this is just equal to $H(X)$. Similarly, the third term is just $H(Y)$. Putting it all together yields

$$D(p_{XY}||p_X \otimes p_Y) = -H(XY) + H(X) + H(Y) \tag{5.64}$$
$$\geq 0 \tag{5.65}$$

where the inequality follows because relative entropy is non-negative. □

As a final application, we will prove that entropy is concave.

Corollary 5.2.3.

Let $\{p_x\}$ be a set of probability distributions, and π_x be a a probability distribution. Then

$$\sum_x \pi_x H(p_x) \leq H(\sum_x \pi_x p_x) \tag{5.66}$$

Proof. Define $p(x,y) = \pi_x p_x(y)$. Then

$$H(Y|X) = H(X,Y) - H(X) \tag{5.67}$$
$$= -\sum_{x,y} \pi_x p_x(y)\log(\pi_x p_x(y)) - \sum_x \pi_x \log(1/\pi_x) \tag{5.68}$$
$$= \sum_x \pi_x H(p_x) \tag{5.69}$$

and

$$H(Y) = H\left(\sum_x \pi_x p_x\right). \tag{5.70}$$

Concavity then follows from the fact that $H(Y|X) \leq H(Y)$. □

5.2.1 Hypothesis Testing

Hypothesis testing is concerned with the following question: suppose we have two distributions p and q, and we get a sample x that we are told came either from p or q. How can we decide which?

There are two possible mistakes you could make, which are very descriptively called "type 1 error" and "type 2 error".

- Type 1 error: You guess $x \sim q$ when actually $x \sim p$. We will use α to denote the probability of a type 1 error.
- Type 2 error: You guess $x \sim p$ when actually $x \sim q$. We will use β to denote the probability of a type 2 error.

There are a few different kinds of hypothesis testing:

- Symmetric hypothesis testing: Come up with a test that minimizes $(\alpha + \beta)/2$, which has a minimum of $\frac{1}{2}||p - q||_1$.
- Bayesian hypothesis testing: Come up with a test that minimizes $\pi\alpha + (1 - \pi)\beta$, which has a minimum of $||\pi p - (1 - \pi)q||_1 + f(\pi)$ for some function f (π is prior probability that it's p.)
- Asymmetric hypothesis testing: Minimize β, subject to the contraint that $alpha < \epsilon$.

  We are going to study asymmetric hypothesis testing. Let $\beta_\epsilon = \min\{\beta|\alpha \leq \epsilon\}$, and $\beta_\epsilon^n = \beta_{epsilon}$ for distinguishing p^n vs q^n (i.e. you get n samples to distinguish p and q).

  As a first observation, note that as we increase n, we should get more confident, and β_ϵ should go decrease. One might hope that it will scale as e^{-nR} for some R, and this indeed turns out to be the case, with R being the relative entropy.

  Theorem 5.2.3 (Chernoff-Stein)**.** *For all $\epsilon \in (0, 1)$,*

$$\lim_{n\to\infty} \frac{-1}{n} \log \beta_\epsilon^n = D(p||q) \tag{5.71}$$

  Instead of proving this theorem, we will look at a few examples to see the sorts of tests that yield the desired β_ϵ.

  Examples

  1. Suppose $p = q$. Then $\Leftrightarrow D(p||q) = 0$ and β_ϵ^n stays constant, as the distributions are identical and there's nothing that can distinguish them.
  2. Suppose q is the uniform distribution. Then $D(p||q) = \log d - H(p)$. Here is a test: take a sample and check if $x^n \in T_{p,\delta}^n$. If it is, guess p, otherwise, guess q. We see that

the probability of a type one error is the probability that the sample is not in $T^n_{p,\delta}$, i.e. $\alpha = p^n(\overline{T^n_{p,\delta}})$ which goes to zero as n goes to infinity. We can also see that

$$\begin{aligned}
\beta &= u^n(T^n_{p\delta}) && (5.72)\\
&= \frac{|T^n_{p,\delta}|}{d^n} && (5.73)\\
&\leq \exp\left(nH(p) + n\delta - n\log d\right) && (5.74)\\
&= \exp\left(-n(D(p||u) - \delta)\right) && (5.75)
\end{aligned}$$

Since we can make δ arbitrarily small, we see that our test obtains the scaling from theorem 5.2.3.

3. Suppose $D(p||q) = \infty$. Then, $A = \text{supp}(p) \setminus \text{supp}(q) \neq \emptyset$, i.e. the support of p is not contained in the support of q. We can then use a simple test: if there exists an x_i in the sample such that $x_i \in A$, then guess p. Otherwise, guess q. Since x_i being in p guarantees that the sample came from p, we get $\beta = 0$, and $\alpha = p(\text{supp }(q))^n \to 0$ as n gets big.

8.372 Quantum Information Science III **Fall 2024**

Lecture 6: September 24, 2024

Scribe: John Blue *Quantum Relative Entropy*

6.1 Chernoff-Stein Lemma

Let's start by proving the Chernoff-Stein Lemma from the last lecture. The setup: we have a string x^n, which was sampled either from p^n or q^n, and we want to know which. To do this, we will look at the likelihood ratio test (LRT). To perform this test, we first compute

$$W(x^n) = \log \frac{p^n(x^n)}{q^n(x^n)}. \tag{6.76}$$

Note that W is a random variable, and that

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

Then, to make the decision, we define some value T such that if $W \geq T$, we guess p^n, and if $W < T$, we guess q^n. Let $A = \{x^n | W(x^n) \geq T\}$ be the "acceptance region".

We're interested in asymmetric hypothesis testing: we need $p^n(A) \geq 1 - \epsilon$ (i.e. the probability that we guess q when it was actually p should be less than ϵ), and then $q^n(A) \leq e^{-nR}$ for some R (the probability that we guess p when it was actually q should grow exponentially small with n).

To decide where to set T, observe that if you set the threshold above $nD(p||q)$, then (in the limit of large sample sizes), we will never guess p. On the other hand, if we set the threshold below $-nD(q||p)$, we will never guess q. This suggests we should set T somewhere inside this range. Since we want to minimize $q^n(A)$, we'll pick T to be closer to this upper bound: $T = n(D(p||q) - \delta)$.

We will first show that this T achieves the desired bound for $p^n(A)$. Consider that

$$p^n(A) = \Pr_{x^n \sim p^n}\left[\log \frac{p^n(x^n)}{q^n(x^n)} > nD(p||q)\right] \tag{6.77}$$

$$= \Pr_{x^n \sim p^n}\left[D(p||q) - \frac{1}{n}\sum_{i=1}^{n} W[x_i] < \delta\right] \tag{6.78}$$

Since $\mathbb{E}_{x \sim p}[W[x]] = D(p||q)$, and each of the x_i are independent and identically drawn from the source, by the law of large numbers, this quantity approaches 1 as n goes to infinity. Thus, for any ϵ and δ we can take n large enough that $p^n(A) \geq 1 - \epsilon$.

Now to show that $q^n(A)$ is small. If $x^n \in A$, then $q^n(x^n) \leq e^{-T} p^n(x^n)$. Then,

$$q^n(A) \leq e^{-T} p^n(A) \tag{6.79}$$

$$\leq e^{-T} \tag{6.80}$$

so $R = D(p||q) - \delta$ (for any $\delta > 0$).

6.1.1 Multiple hypothesis testing

We briefly mention another form of hypothesis testing: multiple hypothesis testing. Here, we have $Q \subseteq \Delta_d = \{\text{prob dists on } [d]\}$. Now, we want to distinguish between the two cases $x^n \sim p^n$ or q^n for some $q \in Q$. Intuitively, it makes sense that distinguishing p from Q should be at least as hard as distinguishing p from q^*, where q^* is the distribution in Q closest to p (see figure 6.2). It turns out that it is actually equally as hard - you can distinguish with the exponential rate

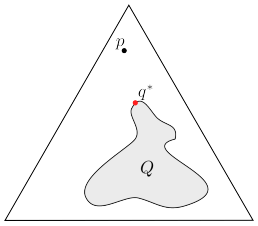


Figure 6.2: An example of multiple hypothesis testing. Given a sample x^n, we want to distinguish between two cases, $x^n \sim p^n$, or $x^n \sim q^n$ where $q \in Q$, a subset of the probability simplex. Here, q^* is the point in Q closest to p.

$R = \min_{q \in Q} D(p||q)$.

6.2 Quantum Relative Entropy and Quantum Chernoff-Stein

We will now turn to the quantum analogue of hypothesis testing. First, we define the quantum relative entropy.

Definition 6.2.1 (Quantum Relative Entropy)**.** *The quantum relative entropy, $D(\rho||\sigma)$, is defined as*

$$D(\rho||\sigma) = Tr[\rho(\log \rho - \log \sigma)].$$

Note that if $[\rho, \sigma] = 0$, this reduces to the classical relative entropy. Just like in the classical case, $D(\rho||\sigma) \geq 0$. From this, we get that

- $S(\rho) \leq d$
- $I(A; B) \geq 0$
- $S(A) \geq S(A|B)$

  We also have a Quantum Pinsker's Inequality.

  Theorem 6.2.1.

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

Now for asymmetric hypothesis testing. Our distributions now will be two quantum states, ρ and σ, and the test will be a set of measurement operators $\{M, I - M\}$ where an outcome of M means we say the state is ρ, and an outcome of $I - M$ means we say the state is σ. We now want to find

$$\beta_\epsilon^n = \min \left\{ \text{Tr}\left[M\sigma^{\otimes n}\right] \mid \text{Tr} M\rho^{\otimes n} \geq 1 - \epsilon \right\}.$$

Theorem 6.2.2 (Quantum Chernoff-Stein Theorem)**.**

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

Before looking at the proof, we will examine the case when ρ and σ are pure and $D(\rho||\sigma) = \infty$, i.e. $\text{supp}(\rho) \nsubseteq \text{supp}(\sigma)$. Let $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$. The measurement that achieves the desired rate is $M = I - B^{\otimes n}$, where $A = |\phi^\perp\rangle\langle\phi^\perp|$, and $B = I - A$. (A result of M is saying that you measured $|\phi^\perp\rangle$ at least once).

Then

$$\text{Tr}(M\sigma^{\otimes n}) = 0$$

while $\text{Tr}(M\rho^{\otimes n}) \to 1$ as $n \to \infty$ (in every register you get some probability of $|\phi^\perp\rangle$, so as $n \to \infty$ you are increasing your chances)

We will now prove the theorem.

Proof. We want an M such that $\text{Tr}(\rho^n M) \geq \alpha$ and $\text{Tr}(\sigma^n M) \leq e^{-nR}$. The idea will be to construct something similar to the LRT, but we will have to be careful about eigenbases.

Let $\rho = \sum_x r_x |\alpha_x\rangle\langle\alpha_x|$ and $\sigma = \sum_x s_x |\beta_x\rangle\langle\beta_x|$.

Recall the definition of a typical projector:

$$\Pi_{p,\delta}^n = \sum_{x^n : |\frac{1}{n}\sum_{i=1}^n \log r_{x_i} + \text{Tr}(\rho\log\rho)| \leq \delta} |\alpha_{x^n}\rangle\langle\alpha_{x^n}|.$$

Next, define

$$\Pi_{\rho||\sigma,\delta}^n = \sum_{x^n : |\frac{1}{n}\sum_{i=1}^n \log s_{x_i} - \text{Tr}(\rho\log\sigma)| \leq \delta} |\beta_{x^n}\rangle\langle\beta_{x^n}|.$$

We note that both of the subspaces defined by these projectors are typical under ρ, *i.e.*, $\text{Tr}(\rho^n \Pi_{p,\delta}^n) \geq 1-\epsilon$ and $\text{Tr}(\rho^{\otimes n}\Pi_{\rho||\sigma,\delta}^n) \geq 1-\epsilon$. We also have that $[\Pi_{\rho,\delta}^n, \rho^{\otimes n}] = 0$ and $[\Pi_{\rho||\sigma,\delta}^n, \sigma^{\otimes n}] = 0$.

If we sandwich $\rho^{\otimes n}$ between the typical projectors, we cut off the "atypical" eigenvalues:

$$e^{-n(S(\rho)+\delta)}\Pi_{p,\delta}^n \leq \Pi_{\rho,\delta}^n \rho^{\otimes n} \Pi_{\rho,\delta}^n \leq e^{-n(S(\rho)-\delta)}\Pi_{\rho,\delta}^n.$$

Similarly, if you do the conditional projection, it squishes the eigenvalues of σ into the following range:

$$e^{n(\text{Tr}(\rho\log\sigma)-\delta)}\Pi_{\rho||\sigma} \leq \Pi_{\rho||\sigma}\sigma^n\Pi_{\rho||\sigma} \leq e^{n(\text{Tr}(\rho\log\sigma)+\delta)}\Pi_{\rho||\sigma}. \tag{6.81}$$

(Note that from here on we will drop the δ and n on the typical projectors).

To get some intuition for equation 6.81, suppose you measure $\log\sigma = \sum_x \log s_x \beta_x$ on ρ. Then

$$\Pr[\log s_x] = \text{Tr}[\rho\beta_x],$$

and the expectation is $\text{Tr}\rho\log\sigma$. If you do this n times, the law of large numbers says that the average will approach the expectation.

We will first show achievability. Our measurement will be the product of both projectors - first measure $\{\Pi_{\rho||\sigma}, I - \Pi_{\rho||\sigma}\}$, and if you get the positive outcome $\Pi_{\rho||\sigma}$, then measure $\{\Pi_\rho, I - \Pi_\rho\}$.

More rigorously, define

$$M = \Pi_{\rho||\sigma}\Pi_\rho\Pi_{\rho||\sigma}.$$

Then

$$\text{Tr}\left(\rho^{\otimes n} M\right) = \text{Tr}\left(\Pi_\rho\Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}\Pi_\rho\right).$$

We need to show that the probability of ρ accepting is large. As we saw above, the probability of ρ accepting for each individual measurement is large. To show the combination works, we can use the Gentle Measurement lemma, which says that the state after accepting $\Pi_{\rho||\sigma}$ is still very close to ρ:

$$||\rho^{\otimes n} - \Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}||_1 \leq 2\sqrt{\epsilon}.$$

We then see that

$$\mathrm{Tr}(\Pi_\rho(\rho^{\otimes n} - \Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma})) \leq \frac{1}{2}||\rho^{\otimes n} - \Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}||_1 \leq \sqrt{\epsilon}$$

from which it follows that

$$\mathrm{Tr}(\Pi_\rho\Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}\Pi_\rho) \geq \mathrm{Tr}(\Pi_\rho\rho^{\otimes n}) - \sqrt{\epsilon} \geq 1 - \epsilon - \sqrt{\epsilon}.$$

So we have now showed that type one error is small enough, and now we need to bound type two error. Consider that

$$\begin{aligned}
\mathrm{Tr}(M\sigma^{\otimes n}) &= \mathrm{Tr}\left(\Pi_\rho\Pi_{\rho||\sigma}\sigma^{\otimes n}\Pi_{\rho||\sigma}\right) && (6.82)\\
&\leq \mathrm{Tr}\left(\Pi_\rho e^{n(\mathrm{Tr}(\rho\log\sigma)+\delta)}\Pi_{\rho||\sigma}\right) && (6.83)\\
&\leq e^{n(S(\rho)+\delta)}e^{n(\mathrm{Tr}(\rho\log\sigma)+\delta)} && (6.84)\\
&= e^{-n(D(\rho||\sigma)-2\delta)} && (6.85)
\end{aligned}$$

where we used the operator inequality from equation 6.81, the fact that $\Pi_{\rho||\sigma} \leq I$, and that $\mathrm{Tr}(\Pi_\rho) = |T_p| \leq e^{n(S(\rho)+\delta)}$.

We have now shown achievability. Next, we will use similar arguments to show the converse, i.e., you cannot do better than the rate $D(\rho||\sigma)$.

Suppose $\mathrm{Tr}\left(M\rho^{\otimes n}\right) \geq \alpha$. Our goal is to show that $\mathrm{Tr}\left(M\sigma^{\otimes n}\right)$ is "not too small". We will use the following operator inequalities:

$$\begin{aligned}
\sigma^{\otimes n} &\geq \Pi_{\rho||\sigma}e^{n(\mathrm{Tr}(\rho\log\sigma)-\delta)} && (6.86)\\
\Pi_\rho\rho^{\otimes n}\Pi_\rho &\leq e^{-n(S(\rho)-\delta)}\Pi_\rho. && (6.87)
\end{aligned}$$

Now

$$\begin{aligned}
\mathrm{Tr}\left(M\sigma^{\otimes n}\right) &\geq \mathrm{Tr}\left(\Pi_{\rho||\sigma}M\right)e^{n(\mathrm{Tr}(\rho\log\sigma)-\delta)} && (6.88)\\
&\geq (\alpha - \sqrt{2\epsilon})e^{-n(D(\rho||\sigma)-2\delta)} && (6.89)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Tr}\left(\Pi_{\rho||\sigma}M\right) &\geq \mathrm{Tr}\left(\Pi_{\rho||\sigma}M\Pi_{\rho||\sigma}\Pi_\rho\right) && (6.90)\\
&\geq \mathrm{Tr}\left(M\Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}e^{-n(S(\rho)-\delta)}\right). && (6.91)
\end{aligned}$$

We can again use Gentle Measurement to show that $\Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}$ is close to $\rho^{\otimes n}$, and using a similar argument as before, we get that

$$\mathrm{Tr}\left(M\Pi_{\rho||\sigma}\rho^{\otimes n}\Pi_{\rho||\sigma}\right) \geq \alpha - \sqrt{2\epsilon}.$$

Putting it all together, we find

$$\mathrm{Tr}\left(M\sigma^{\otimes n}\right) \geq (\alpha - \sqrt{2\epsilon})e^{-n(D(\rho||\sigma)-\delta)} \tag{6.92}$$

which completes the proof. □

8.372 Quantum Information Science III **Fall 2024**

Lecture 7: September 26, 2024

Scribe: Jonathan Lu *Noisy channel coding*

7.1 Aside: concavity of quantum entropy

Suppose we have two density matrices ρ_0 and ρ_1. We can mix them together with some probability weight π to obtain $\rho := \pi\rho_0 + (1-\pi)\rho_1$. The concavity property of the quantum entropy S tells us that $S(\rho)$ is at least as large as the mixed entropy $\pi S(\rho_0) + S(\rho_1)$. Because it's so important, let's prove the concavity of S.

Theorem 7.1.1. $S(\rho) = S(\pi\rho_0 + (1-\pi)\rho_1) \geq \pi S(\rho_0) + (1-\pi)S(\rho_1)$.

Proof. Let $\sigma^{AB} = \pi\rho_0^A \otimes |0\rangle\langle 0|^B + (1-\pi)\rho_1^A \otimes |1\rangle\langle 1|^B$. This is the *labeled* mixture, so that if we have access to the B system we know which density matrix we have. Note now that

$$S(A) = S(\rho),\ S(B) = H_2(\pi) := -\pi\log\pi - (1-\pi)\log(1-\pi). \tag{7.93}$$

Also, by definition, $S(A|B) = S(AB) - S(B)$ and $S(AB) = -\operatorname{tr}[\sigma\log\sigma]$. The structure of σ makes it block diagonal, since

$$\sigma = \left(\begin{array}{c|c} \pi\rho_0 & 0 \\ \hline 0 & (1-\pi)\rho_1 \end{array}\right), \quad \log\sigma = \left(\begin{array}{c|c} \log\rho_0 + (\log\pi)I & 0 \\ \hline 0 & \log\rho_1 + \log(1-\pi)I \end{array}\right). \tag{7.94}$$

This block diagonal structure makes the calculation of the joint entropy simple:

$$S(AB) = -\operatorname{tr}[\sigma\log\sigma] = -\pi\operatorname{tr}[\rho_0\log\rho_0] - \pi\log\pi - (1-\pi)\operatorname{tr}[\rho_1\log\rho_1] - (1-\pi)\log(1-\pi) \tag{7.95}$$

$$= H_2(\pi) + \pi S(\rho_0) + (1-\pi)S(\rho_1) \tag{7.96}$$

$$= S(B) + S(A|B). \tag{7.97}$$

We observe that the conditional entropy takes a simple form because the system being conditioned upon is just a classical probability distribution:

$$S(A|B) = \pi S(\rho_0) + (1-\pi)S(\rho_1). \tag{7.98}$$

Therefore, $S(\rho) - [\pi S(\rho_0) + (1-\pi)S(\rho_1)] = S(A) - S(A|B) = I(A;B) \geq 0$. The last inequality follows from the fact that $I(A;B) = D(\rho_{AB}||\rho_A \otimes \rho_B) \geq 0$, as we saw in the classical case. The proof that quantum relative entropy is non-negative is delegated to Problem Set 3. □

7.2 Classical noisy channel coding

In lecture 3, we stated Shannon's noisy coding theorem. Today we will prove it. Recall that a *channel* is a conditional probability distribution $N(y|x)$, so that if your input source is the distribution $\pi(x)$, the joint distribution of input-output pairs is $p(x,y) = \pi(x)N(y|x)$. The *capacity* of a channel is defined to encode the most amount of information you can send through the channel with asymptotically small noise.

Definition 7.2.1. *For a channel N, the capacity is given by*

$$C(N) = \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log |M(\epsilon, N)|, \tag{7.99}$$

where $M(\epsilon, N)$ is the set of messages that can be sent through N^n with error probability $\leq \epsilon$.

Figure 7.3 shows the model we will adopt, in which we encode a set of messages M into bits before it is sent through a noisy channel, after which the noisy message is decoded into something that is hopefully the original message.

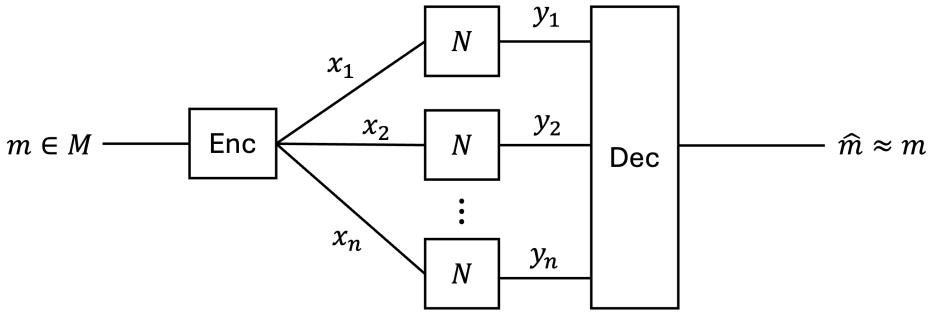


Figure 7.3: Encoder-decoder model with a channel in bewteen them.

Theorem 7.2.1 (Shannon's noisy coding theorem)**.** $C(N) = \max_\pi I(X;Y)$.

Before we prove the theorem, let's assume it's true and look at some illustrative examples of channels. For just these examples, let π be the probability that $X = 0$; we will only do examples with a single bit.

1. Binary symmetric channel with error probability η. Let x, y be single bits. Then $y = x \oplus \rho$, where $\Pr[\rho = 1] = \eta$. So, the bit gets flipped with probability η. Then

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_2(\eta). \tag{7.100}$$

   To maximize, note that $H_2(\eta)$ does not depend on π, and $H(Y) \leq 1$. But if $\pi = 1/2$, then $H(Y) = 1$. Hence for any δ, we can asymptotically send $n(1 - H_2(\eta) - \delta)$ bits of information to the output using n bits of input.

2. Erasure channel. Regardless of the input bit, there is a probability η that the channel maps it to $\perp$ (the "erased" message). Then $H(Y|X) = H_2(\eta)$ as with the binary symmetric channel. To calculate $H(Y)$, we note that $Y \in \{0, 1, \perp\}$ with probabilities $\pi(1-\eta), (1-\pi)(1-\eta), \eta$. Therefore,

$$H(Y) = -\pi(1-\eta)\log[\pi(1-\eta)] - (1-\pi)(1-\eta)\log[(1-\pi)(1-\eta)] - \eta \log \eta \tag{7.101}$$
$$= H_2(\eta) + (1-\eta)H_2(\pi). \tag{7.102}$$

   So we want to maximize $I(X;Y) = H(Y) - H(Y|X) = (1-\eta)H_2(\pi)$, which occurs when again $\pi = 1/2$, giving a capacity $C = 1 - \eta$.

The erasure channel capacity result is particularly remarkable. Consider a situation in which you and your friend are talking over the phone. Sometimes, the phone glitches with probability η and

erases whatever your friend said at that time. To remedy this, you might say "what?", asking your friend to repeat herself. This protocol has an obvious capacity of $1-\eta$, but it also involves *feedback*, allowing the receiver to send information back to the sender. Shannon's theorem implies by the above that *even without feedback*, you can achieve the same capacity!

Now we want to actually prove Theorem 7.2.1. Let Alice send input bits and Bob receive output bits. Alice will send bits $x_1, \dots, x_n$ and Bob receives $y_1, \dots, y_n$. The intuition for this proof is that we will only worry about the typical set of Y^n, of which there are about $2^{nH(Y)}$, since those are the only strings that will be sent asymptotically. On the other hand, for a given input string x^n, there are about $2^{nH(Y|X)}$ output strings y^n that could have reasonably come from x^n. To ensure decodability, we want these possible string sets for each distinct (typical) x^n not to overlap. That implies we can have at most $2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(I;Y)}$ codewords.

Today we prove the achievability portion of the theorem, and leave the converse to next time.

Lemma 7.2.1. *$C(N) \geq \max_\pi I(I;Y)$. That is, for any rate $R < \max_\pi I(I;Y)$, there exists an encoding procedure that decodes with asymptotically vanishing error probability.*

Proof. For the proof, we'll switch back to $\pi = \pi(x)$ being a distribution over x. We formalize the above intuition by using relative entropy. Define $q_x(y) = N(y|x)$ just for notation and let $q(y) = \sum_x \pi(x) q_x(y)$ be the marginal distribution on Y. Then

$$D(q_x||q) = \sum_y q_x(y) \log \frac{q_x(y)}{q(y)} = -H(q_x) - \sum_y q_x(y) \log q(y). \tag{7.103}$$

The relation between relative entropy of these distributions and mutual information becomes clear when we sum over x:

$$\sum_x \pi(x) D(q_x||q) = -\sum_x \pi(x) H(q_x) - \sum_{x,y} \pi(x) q_x(y) \log q(y) \tag{7.104}$$

$$= -H(Y|X) + H(Y) = I(X;Y). \tag{7.105}$$

Define $x^n(m) := \mathrm{Enc}(m)$ and consider only $x^n(m) \in T^n_\pi$ the typical space, i.e. where i appears $n\pi_i$ times. Then $N^n(x^n(m)) = q_{x_1} \otimes \cdots \otimes q_{x_n}$, which up to permutation is $q_1^{\otimes n\pi_1} \otimes \cdots \otimes q_d^{\otimes n\pi_d}$. Note that $N^n(x^n(m))$ is itself a probability function and can be evaluated on strings and sets of strings, so to avoid confusion we will write $N^n(x^n(m))[S]$ as the conditional probability of getting strings in S as output given $x^n(m)$ as input. Since relative entropies add for independent distributions,

$$D(N^n(x^n(m)), q^{\otimes n}) = nI(X;Y). \tag{7.106}$$

By Stein's lemma from last lecture, there exists for any choices of ϵ, δ, a test set $A(m) \subseteq [d]^n$ such that $N^n(x^n(m))[A(m)] \geq 1-\epsilon$ but $q^n(A(m)) \leq 2^{-n(I(X;Y)-\delta)}$.

With these guarantees, we are ready to write down our encoding and decoding procedures. For the encoding, for each $m \in M$, Alice chooses $x^n(m) \in T_\pi$ (or, nearly equivalently, randomly from π^n). Note that the marginal probability holds as expected:

$$\mathbb{E}_{x^n(m) \sim \pi^n} N^n(x^n(m)) \approx q^{\otimes n}. \tag{7.107}$$

To decode, Bob brute-force iterates through $A(m)$, $m \in M$ and outputs m when the test passes. The probability the test fails is

$$\Pr[\text{error}] = \Pr[\text{wrong test accepts}] + \Pr[\text{right test rejects}] \tag{7.108}$$

$$\leq |M| 2^{-n(I(X;Y)-\delta)} + \epsilon. \tag{7.109}$$

By construction, $|M| = 2^{nR}$. Thus, if $R < I(X;Y) - \delta$, the first term asymptotically vanishes and so the error probability will asymptotically be ϵ, as desired. □

8.372 Quantum Information Science III **Fall 2024**

Lecture 8: October 1st 2024

Scribe: Adam Wills and Daniel Lee *Classical Messages over Quantum Channels*

8.0 Lecture Outline

The aim of today's lecture is to cover three things. First, we'd like to establish a quantum version of Shannon's noisy channel coding theorem that we discussed last time. In particular, we'd like to discuss sending classical messages over quantum channels, and establish the rate at which we can send information in such a situation. Having established the achievability of this, in a fairly analogous way to what we did last time for Shannon's noisy channel coding theorem, we will turn to the converse for both Shannon's noisy channel coding theorem and for this quantum situation. We won't have time to prove this completely, but as preparation for the upcoming proof, we will introduce the Conditional Mutual Information (CMI).

8.1 Classical-Quantum (CQ) Channels

We will talk about channels with classical input and quantum output, otherwise known as CQ channels; for example,

$$\mathcal{N} : x \mapsto \rho_x. \tag{8.110}$$

These can be imagined as special cases of usual CPTP quantum channels, for example,

$$\mathcal{N}(|x\rangle \langle y|) = \delta_{xy}\rho_x. \tag{8.111}$$

This also corresponds to a quantum channel which measures its input, before sending on some quantum states dependent on the outcome of this measurement. By specialising to classical-quantum channels, we avoid many of the difficulties experienced with general quantum-quantum channels, represented by a general CPTP map, for which entangled inputs are allowed, and many results become very hard to prove.

8.1.1 The HSW Theorem

The HSW Theorem tells us the capacity of such a channel. It is

$$C(\mathcal{N}) = \max_p I(X : Q)_\omega, \tag{8.112}$$

where the maximum is taken over all probability distributions p, and ω is the "classical-quantum state"

$$\omega^{XQ} = \sum_x p(x) |x\rangle \langle x|^X \otimes \rho_x^Q. \tag{8.113}$$

Comments:

1. As a special case, if ρ_x is diagonal, then this reduces to Shannon's noisy channel coding theorem.

2. We have $I(X:Q) = S(Q) - S(Q|X)$ as always. It's worth appreciating that asking for the entropy of Q conditioned on X makes more sense for this state than for a general quantum state; conditioning on the outcome of some classical random variable is much more meaningful than conditioning on some quantum state. Notice that

$$I(X:Q)_\omega = S(\rho) - \sum_x p(x) S(\rho_x) \tag{8.114}$$

where $\rho = \sum_x p(x)\rho_x$ is the average state. It is common to denote a quantity called Holevo's χ-quantity as

$$\chi(\mathcal{N}) = \max_p \left[S(\rho) - \sum_x p(x) S(\rho_x) \right], \tag{8.115}$$

so that $C(\mathcal{N}) = \chi(\mathcal{N})$.

3. As is typically the case, the theorem comes with an achievability part and a converse, so a full proof shows that information transmission at a rate of $C(\mathcal{N})$ is possible, and that attempting to transmit information at any faster rate will fail (i.e. the error rate in communication will become large).

4. The HSW theorem has an interesting relation to the scenario of *accessible information*. Suppose Alice wants to send classical information to Bob via a quantum channel. She wishes to encode some input x, corresponding to the value of some random variable $X \sim p$, into some quantum state ρ_x which then gets sent to Bob. Bob then tries to learn about x by performing some measurement $\{M_y\}_y$ and deducing the outcome y as a result, with the hope that $y = x$.

   This whole thing can be considered as a classical channel, and the corresponding mutual information (maximised over the best possible measurement by Bob) is known as the accessible information of the ensemble $\{p_x, \rho_x\}$:

$$I_{acc}(\{p_x, \rho_x\}) = \sup_{\{M_y\}_y} I(X:Y). \tag{8.116}$$

   It is true in general that

$$I_{acc}(\{p_x, \rho_x\}) \leq S(\rho) - \sum_x p(x) S(\rho_x). \tag{8.117}$$

Example 8.1.1. *Consider a C-Q channel that sends the classical input i to the quantum output ρ_i for $i = 1, 2, 3$. Suppose the ρ_i are pure qubit states lying in the equator of the Bloch sphere (see Figure 8.4), and the three of them are mutually equally spaced. In this case, we have that the average state ρ is the maximally mixed state, and so $S(\rho) = 1$. We also have that the entropy of each individual state is 0 (because the states are pure). As such, we have $C(\mathcal{N}) = 1$. We might consider this to be somehow surprising, because given only one transmission of ρ_1, ρ_2, ρ_3, it is impossible to reliably distinguish between them. However, the definition of $C(\mathcal{N})$ is an asymptotic statement. We are seeing that asymptotically, $\mathcal{N}$ is as good as a classical noiseless channel just transmitting one bit.*

This is already giving us the indication that to get the most out of this communication scenario, the receiver, Bob, must perform entangled measurements on the outputs of the n uses of the channel

$$\rho_{x_1} \otimes \rho_{x_2} \otimes \ldots \otimes \rho_{x_n}. \tag{8.118}$$

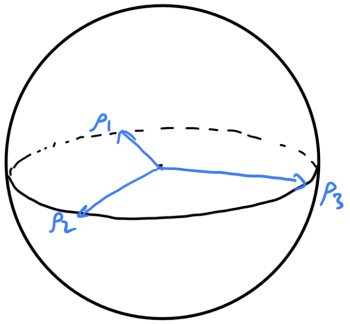


Figure 8.4: Three pure states on the equator of the Bloch sphere, which average to the maximally mixed state

8.2 Quantum-Quantum Channels

We will quite shortly go on to prove the achievability portion of the HSW theorem, and then direct ourselves towards the converse. First, however, let us make some comments on general quantum-quantum channels, which are in general harder to deal with.

A general quantum-quantum channel is represented by a CPTP map $\mathcal{N}$, and in this case we define the χ-quantity as

$$\chi(\mathcal{N}) = \max_{\{p_x, \sigma_x\}} I(X:Q)_\omega, \tag{8.119}$$

where

$$\omega = \sum_x p_x \left|x\right\rangle \left\langle x\right| \otimes \mathcal{N}(\sigma_x). \tag{8.120}$$

The capacity of the channel is then in fact

$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} \chi(\mathcal{N}^n). \tag{8.121}$$

It is an interesting and very much non-trivial fact, known as the superadditivity of quantum channel capacity, that in general one can send more information through a channel $\mathcal{N}$ by sending entangled inputs through multiple uses of it, then just using it separately multiple times, i.e., $C(\mathcal{N}) > \chi(\mathcal{N})$ in general.

Comments:

1. The above formula for quantum channel capacity is very hard to work with in practice. To see this, let us start by considering the formula given by the HSW theorem. This is a relatively easy expression to work with, because it is a maximisation of a concave function ($I(X:Q)_\omega$) over a convex set (the set of probability distributions). This makes it easy to prove things about in theory, and also our computers can handle it easily with standard optimisation techniques.

   Conversely, the general quantum channel capacity is not a maximisation of a concave function. It is in fact NP-hard in general to computationally find the maximising solution of $\chi(\mathcal{N})$, so we do not expect to find efficient techniques for performing this optimisation.

2. By applying this formula to a noiseless qubit channel, one can see that it is impossible to reliably send more than n bits of information in n qubits, despite the fact that a full description of an n-qubit state requires 2^n complex numbers. Qubits are therefore no better carriers of information than classical bits, even if they are better at certain other tasks, like secret sharing.

8.3 Proof of Achievability for the HSW Theorem

To prove the achievability part of the HSW theorem, we want another tool.

8.3.1 Non-Commutative Union Bound

Classically, suppose we have some bad events that are unlikely to happen. Via the standard union bound, it is easy to say that the probability that at least one bad event happens is the sum of the probabilities of all the individual bad events.

Quantumly, the analogous question might be to consider some density matrix ρ and some two-outcome measurements, each corresponding to operators $\{P_i, I - P_i\}$ for $i = 1, \ldots, l$. Supposing that each P_i is fairly likely to be measured, what is the probability of measuring all $P_1, \ldots, P_l$ on ρ sequentially? We cannot just use the classical case because the state gets disturbed with each measurement. The statement of the lemma is this.

Lemma 8.3.1. *If $\rho \geq 0$, $Tr(\rho) \leq 1$, and $P_1, \ldots, P_l$ are projectors, then*

$$Tr(\rho) - Tr(P_1, \ldots P_l \rho P_l \ldots P_1] \leq 2\sqrt{\sum_{i=1}^{l} Tr((I - P_i)\rho)} \tag{8.122}$$

8.3.2 Remainder of the Proof

Let p achieve the maximum in

$$C(\mathcal{N}) = \max_p (I(X : Q))_\omega. \tag{8.123}$$

Just as in the classical case, let us take a random codebook, so Alice chooses some codewords

$$X^n(1), \ldots, X^n(M) \sim p^n \tag{8.124}$$

identically and independently. For each $m \in [M]$, the state Bob receives is

$$\sigma_m = \rho_{X_1(m)} \otimes \ldots \otimes \rho_{X_n(m)}, \tag{8.125}$$

where $X_i(m)$ is the i-th symbol in the codeword corresponding to $m \in [M]$. Notice that if you average over the choice of codeword $X^n(m)$, you get

$$\mathbb{E}_{X^n(m)} \sigma_m = \left(\sum_x p(x) \rho_x \right)^{\otimes n} = \rho^{\otimes n}. \tag{8.126}$$

Let us define the conditionally typical projector Π_m, for σ_m. Letting t be the type of $X^n(m)$[4], Π_m is defined by

$$\Pi_m = \Pi \bigotimes_{i=1}^{d} \Pi_{\rho_i, \delta}^{n t_i} \Pi^{-1}, \tag{8.127}$$

[4]It is worth appreciating that the type of $X^n(m)$ will be very close to p (for large n).

where Π is the permutation mapping the states ρ_i as they appear in ascending order corresponding to the type, to the order in which they appear in σ_m, i.e. it maps

$$\bigotimes_{i=1}^{d} \rho_i^{nt_i} \mapsto \bigotimes_{i=1}^{n} \rho_{X_i(m)} = \sigma_m. \tag{8.128}$$

Since Π_m is a typical projector for σ_m, we have that

$$Tr(\Pi_m \sigma_m) \geq 1 - \epsilon \tag{8.129}$$

for each m.

In complete analogy to Bob's decoding procedure for Shannon's noisy channel coding theorem, Bob will do nothing other than to sequentially measure $\Pi_1, \Pi_2, \ldots$, and accept the first $m \in [M]$ for which the measurement of Π_m succeeds. We know that the chance of making the correct measurement is high, since

$$Tr(\Pi_m \sigma_m) \geq 1 - \epsilon \tag{8.130}$$

for each m. Let us consider the chance of failure. The non-commutative union bound justifies taking an upper bound for the chance of making the wrong measurement on a message m as simply the sum of making the wrong measurement $\Pi_{\hat{m}}$ on σ_m for each $\hat{m} \neq m$ (because we can ignore the factor of two and the square-root, because they asymptotically make no difference to the rate). The expectation (taken over all codebooks) of making the wrong measurement on a message m is

$$\sum_{\hat{m}:\hat{m}\neq m} \mathbb{E}_{X^n(m),X^n(\hat{m})} Tr(\Pi_{\hat{m}} \sigma_m) = \sum_{\hat{m}:\hat{m}\neq m} \mathbb{E}_{X^n(\hat{m})} Tr(\Pi_{\hat{m}} \rho^{\otimes n}) \tag{8.131}$$

$$= \underbrace{(M-1)}_{2^{nR}-1 \approx 2^{nR}} \exp\left(\underbrace{-\sum_{i=1}^{d} nt_i D(\rho_i||\rho)}_{-nI(X:Q)} \right) \tag{8.132}$$

$$\approx 2^{nR} 2^{-nI(X:Q)}, \tag{8.133}$$

so that indeed if $R < I(X:Q)$, then the probability of making the wrong measurement on m goes to zero. We need to, however, justify the claim that

$$-\sum_{i=1}^{d} nt_i D(\rho_i||\rho) \approx -nI(X:Q). \tag{8.134}$$

This is done quite straightforwardly, however, from the relation

$$I(X:Q)_\omega = \sum_i p(i) D(\rho_i||\rho) \tag{8.135}$$

and then the argument

$$|\sum_i nt_i D(\rho_i||\rho) - \sum_i np(i) D(\rho_i||\rho)| \leq n \underbrace{||t-p||}_{\text{Small with high probability}} \underbrace{\max_i D(\rho_i||\rho)}_{\leq \log d}. \tag{8.136}$$

This concludes the proof, although we comment that making this fully rigorous would mean taking care of various details that have been omitted. For example, taking the average over all codebooks

in the probability of making the wrong measurement only means that some codebook would work - we must fix such a codebook. Also, the last statement that $||t-p||$ is small with high probability is true, although we would need to get rid of all the m's for which this is large. By standard arguments that are very similar to that made last time (for the classical case), this does not mean getting rid of too many messages m.

This concludes our discussion of the proof of achievability.

8.4 Towards a Converse

We will not have full time to prove a converse in this lecture but will start to develop the tools to support the subsequent proof.

8.4.1 Fano's Inequality and Fannes' Inequality

If M and $\hat{M}$ are two random variables (over the same alphabet) that are very likely to be equal, then their conditional entropy is small. In particular, suppose that M and $\hat{M}$ are random variables over the set $\{0,1\}^{nR}$. Then

$$\mathbb{P}[M \neq \hat{N}] \leq \epsilon \implies H(M|\hat{N}) \leq \epsilon nR + 1. \tag{8.137}$$

Proof. Let p be the distribution for M given $\hat{M}$. Then, letting η be the probability that M does not equal $\hat{M}$ ($\eta \leq \epsilon$), we have

$$p = (1-\eta)1_{\hat{M}} + \eta q, \tag{8.138}$$

where $1_{\hat{M}}$ is the distribution concentrated on the value of $\hat{M}$, and q is some distribution with $q(\hat{M}) = 0$. From this, we compute the entropy of p as

$$-(1-\eta)\log(1-\eta) - \sum_x \eta q(x)(\log\eta + \log q(x)) = \underbrace{H_2(\eta)}_{\leq 1} + \underbrace{\eta}_{\leq \epsilon}\underbrace{H(q)}_{\leq nR}, \tag{8.139}$$

which concludes the proof. □

A further useful result is that of Fannes' Inequality, which we will not prove. This says that

$$|S(\rho) - S(\sigma)| \leq H_2(\epsilon) + \epsilon \log d, \tag{8.140}$$

where

$$\epsilon = \frac{1}{2}||\rho - \sigma||_1 (= T(\rho,\sigma)). \tag{8.141}$$

Note that this can be interpreted as a continuity statement for the von Neumann entropy S.

8.4.2 Conditional Mutual Information and Markov Chains

Another very useful tool for the proof of the converse will be that of the conditional mutual information (CMI), which relates closely with the theory of Markov Chains. The CMI of two random variables X and Y given the random variable Z is the mutual information between the random variables X and Y averaged over the output of Z being fixed, i.e.,

$$I(X:Y|Z) \coloneqq \sum_z p_Z(z) I(X:Y|Z=z) \tag{8.142}$$

$$= H(X|Z) + H(Y|Z) - H(XY|Z) \tag{8.143}$$

$$= H(XZ) + H(YZ) - H(XYZ) - H(Z). \tag{8.144}$$

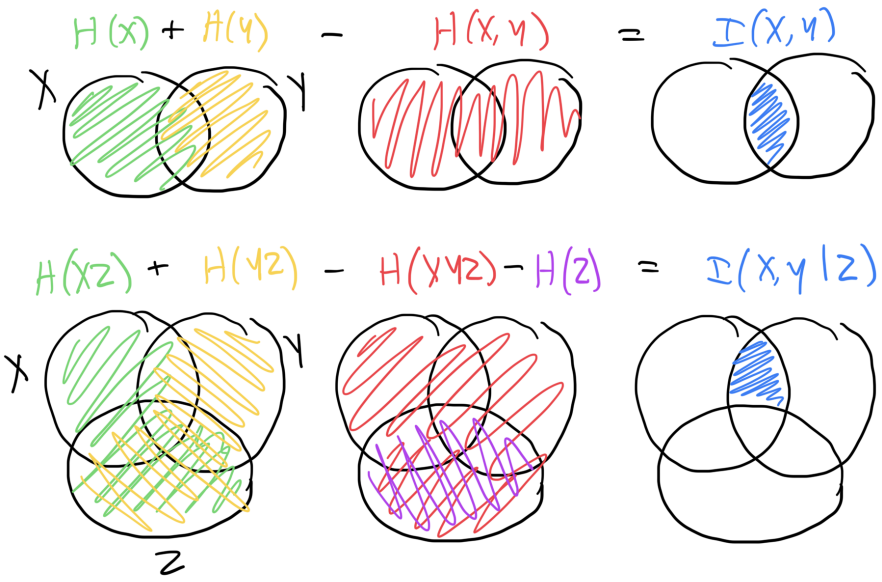


Figure 8.5: Mutual information and conditional mutual information as venn diagrams

These relations can be understood via Venn diagrams (Figure 8.5)

Whereas $I(X:Y)$ may be interpreted as a measure of the correlation between X and Y, $I(X:Y|Z)$ may be interpreted as the correlation between X and Y that remains once you have conditioned on Z. To illustrate this interpretation, we note that $I(X:Y|Z)=0$ if and only if the sequence

$$X \to Z \to Y \tag{8.145}$$

forms a Markov Chain.

8.372 Quantum Information Science III **Fall 2024**

Lecture 9: October 3rd 2024

Scribe: Adam Wills *Converse to Channel Capacity Theorems and Applications*

9.0 Recap from Last Time

Last time we started the proof of the converse of our channel capacity theorems, both for the classical channel capacity (Shannon's noisy channel coding theorem) and for CQ channels (the HSW theorem). We started to make preparations in this direction by stating Fano's inequality, and we started to discuss the conditional mutual information (CMI). We will continue these discussions now to prove the converse.

9.1 Proof of the Converses

Let us start by considering the general encode - channel - decode scenario:

$$M \overset{\text{Encode}}{\longrightarrow} X^n \overset{\text{Channel}}{\longrightarrow} Y^n \overset{\text{Decode}}{\longrightarrow} \hat{M} \tag{9.146}$$

In a successful protocol, we have that the probability M and $\hat{M}$ differ is at most ϵ. Also, as usual, we say that the alphabet from which M is drawn is of size 2^{nR}. We can make a first step in our proof of a converse by applying Fano's inequality. Considering a uniformly random initial message, M, we have $H(M) = nR$. Then, the mutual information between M and $\hat{M}$ is

$$I(M : \hat{M}) = H(M) - H(M|\hat{M}) \geq (1 - \epsilon)nR - 1, \tag{9.147}$$

where we have applied Fano's inequality. We wish to turn this into a statement about the mutual information between X^n and Y^n, and for this we will talk about the CMI.

9.1.1 Conditional Mutual Information (CMI)

Either quantumly or classically, we can define the CMI as

$$I(X : Y|Z) = H(XZ) + H(YZ) - H(XYZ) - H(Z), \tag{9.148}$$

where quantumly the H's are replaced by S's. Classically we can re-write this definition as

$$I(X : Y|Z) = \sum_z I(X : Y|Z = z)p_Z(z), \tag{9.149}$$

although conditioning in this way doesn't make sense quantumly. It's interesting to notice that this expression means that showing the CMI is non-negative classically is no harder than showing the regular mutual information is non-negative, but because this expression isn't meaningful quantumly, one must work a little harder to show non-negativity of CMI quantumly.

With a little rearrangement of entropies, one can quickly see that

$$I(X : Y|Z) = I(X : YZ) - I(X : Z), \tag{9.150}$$

for which the further rearrangement

$$I(X:YZ) = I(X:Y|Z) + I(X:Z) \tag{9.151}$$

gives us the "chain rule" of mutual information. These expressions lend further weight to the interpretation of the CMI $I(X:Y|Z)$ as the amount of information shared by X and Y once you have conditioned on Z. The equation (9.150) tells us that $I(X:Y|Z)$ is the amount of information X knows about YZ that it doesn't already know about Z. Therefore, $I(X:Y|Z)$ is zero if and only if all of the interactions between X and Y are mediated by Z, which says exactly that $X \to Z \to Y$ forms a Markov chain, as stated last time. Classically, this is easy to prove, whereas quantumly we take it as the definition of a quantum Markov chain.

Let us put some more meat on this Markov chains idea in the classical case. We have

$$I(X:Y|Z) = 0 \iff X \to Z \to Y \text{ is a Markov chain} \tag{9.152}$$

$$\iff p(x,y,z) = p(z)p(x|z)p(y|z) = p(x)p(z|x)p(y|z) \tag{9.153}$$

$$\iff p(x,y,z) = f(x,z)g(y,z) \tag{9.154}$$

for some functions f, g. There is a corresponding robustness statement, which is

$$I(X:Y|Z)_p = \min_{q:q \text{ is a Markov Chain}} D(p||q), \tag{9.155}$$

so that if the joint distribution of X, Y and Z is near a Markov chain (in relative entropy), then the CMI is small. This fits into our general family of similar statements:

$$H(X) \approx 0 \iff X \text{ nearly deterministic} \tag{9.156}$$

$$S(\rho) \approx 0 \iff \rho \text{ nearly pure} \tag{9.157}$$

$$H(X|Y) = 0 \iff X \text{ is a deterministic function of } Y \tag{9.158}$$

$$S(X|Y) = 0 \text{ is an exception — no special meaning!} \tag{9.159}$$

$$D(\rho||\sigma) \approx 0 \iff \rho \text{ is almost } \sigma \tag{9.160}$$

We can also provide a physical interpretation to the CMI being zero. Thinking of the random variables as physical systems interacting, we find that if $I(X:Y|Z) = 0$ then the chain $X - Z - Y$ has only local interactions. Thinking in terms of statistical mechanics, their probability distribution factorises into two Gibbs distributions:

$$p(x,y,z) = \frac{e^{-E_1(x,z)-E_2(y,z)}}{Z}. \tag{9.161}$$

This idea extends to more general networks. Suppose we have physical systems X, Z, Y and W interacting locally via the network shown in the Figure. Removing the system Z, or conditioning

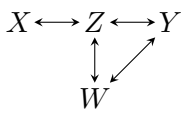


Figure 9.6:

on it, leads to X and W being independent, meaning that the CMI $I(X:W|Z)$ is zero — all interactions between X and W are mediated (in some way) by Z.

Note that not all our discussion of CMI extends to the quantum case. Again, our definition of a quantum Markov chain is simply a state whose CMI is zero, and it is harder to make a robustness statement like Equation (9.155) in the quantum case. Again, non-negativity of CMI is true, but harder to prove. To do so, let us discuss the Data Processing Inequality.

9.1.2 Data Processing Inequality

We will see more on the DPI shortly, but our first form of the DPI will be a classical statement about the Markov chain $X - Z - Y$, for which we have

$$I(X : Z) \geq I(X : Y), \tag{9.162}$$

i.e., the information shared between X and Z is always at least the information shared between X and Y. This is easily proved using the non-negativity of CMI.

Proof.

$$I(X : Z) = I(X : YZ) - I(X : Y|Z) \tag{9.163}$$
$$I(X : Y) = I(X : YZ) - I(X : Z|Y) \tag{9.164}$$

and so

$$I(X : Z) - I(X : Y) = -I(X : Y|Z) + I(X : Z|Y). \tag{9.165}$$

We know the first term is zero using the fact that $X - Z - Y$ forms a Markov chain, and the second term is non-negative, giving the conclusion. □

9.1.3 Strong Subadditivity

Showing that the CMI is non-negative quantumly is harder than classically, and is equivalent to the statement of strong subadditivity, which is the statement that

$$S(XZ) + S(YZ) \geq S(XYZ) + S(Z), \tag{9.166}$$

which is stronger than the usual statement of subadditivity, which recall is

$$S(X) + S(Y) \geq S(XY), \tag{9.167}$$

which itself is equivalent to the regular mutual information being non-negative. The non-negativity of the quantum CMI was initially proved by Lieb and Ruskai in the 70s, which is a different proof to what we show now.

We want to use the fact that, given a quantum operation $\mathcal{E}$,

$$D(\mathcal{E}(\rho)||\mathcal{E}(\sigma)) \leq D(\rho||\sigma). \tag{9.168}$$

This actually follows immediately from our proof of the operational interpretation of $D(\rho||\sigma)$ as the optimal rate at which ρ and σ can be distinguished in an asymmetric hypothesis test. When trying to distinguish ρ and σ, one thing you could do is apply $\mathcal{E}$ to both of them, and then distinguish the resulting states. This immediately gives us this monotonicity statement.

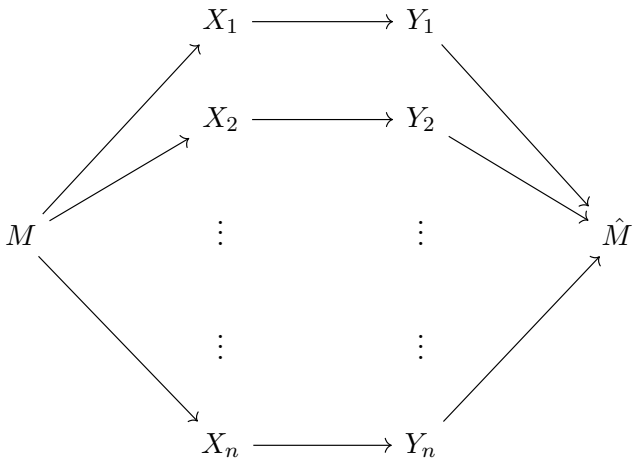
Applying this with $\mathcal{E}$ a partial trace over the system Y, we have

$$I(X : Z) = D(\rho_{XZ}||\rho_X \otimes \rho_Z) \leq D(\rho_{XYZ}||\rho_X \otimes \rho_{YZ}) = I(X : YZ), \tag{9.169}$$

and the result follows.

9.1.4 Back to Channel Capacities

Recall we have the dependence diagram of random variables



which can also be simplified to

$$M \to X^n \to Y^n \to \hat{M}. \tag{9.170}$$

We were up to the point in the proof of the converse where we had $(1-\epsilon)nR - 1 \leq I(M : \hat{M})$, and then using DPI twice gives $I(M : \hat{M}) \leq I(X^n : Y^n)$. We then use

$$I(X^n : Y^n) \leq \sum_{j=1}^{n} I(X_j : Y_j) \tag{9.171}$$

which we will show to be true momentarily. The proof of the converse follows immediately, since we find that we have

$$(1-\epsilon)nR - 1 \leq n \max_p I(X_1 : Y_1)_p, \tag{9.172}$$

so that we find that R is indeed at most the claimed capacity

$$C(\mathcal{N}) = \max_p I(X : Y)_p. \tag{9.173}$$

Let us now prove the claim of Equation (9.171).

Proof.

$$I(X^n : Y^n) = H(Y^n) - H(Y^n|X^n) \tag{9.174}$$

$$\leq \sum_{j=1}^{n} H(Y_j) - H(Y^n|X^n) \tag{9.175}$$

by subadditivity. Then, by a chain rule/telescoping sum,

$$H(Y^n|X^n) = \sum_{j=1}^{n} H(Y_j|X^n Y_1 Y_2 \dots Y_{j-1}). \tag{9.176}$$

Referring to our dependency diagram, we can see that once we condition on X_j, further conditioning via the other $X's$ or any other Y's does not change the distribution of Y_j (this is the definition of such dependence). We therefore have

$$H(Y_j|X^n Y_1 Y_2 \dots Y_{j-1}) = H(Y_j|X_j). \tag{9.177}$$

In total,

$$I(X^n : Y^n) \leq \sum_{j=1}^{n} H(Y_j) - H(Y_j|X_j) = \sum_{j=1}^{n} I(X_j : Y_j). \tag{9.178}$$

□

This concludes our proof of the converse for Shannon's noisy channel coding theorem.

9.1.5 Converse to the HSW Theorem

It turns out that the above argument goes straight through to the case of classical-quantum channels relevant to the HSW theorem, because the inputs are classical. The system Y^n becomes the quantum system Q^n, and we have

$$I(X^n : Q^n) = S(Q^n) - S(Q^n|X^n) \tag{9.179}$$

$$\leq \sum_j S(Q_j) - S(Q_j|S_j) \tag{9.180}$$

$$\leq n\chi, \tag{9.181}$$

where $\chi = \max_p I(X : Q)_\omega$, and we recall the classical-quantum state

$$\omega^{XQ} = \sum_x p(x) \left|x\right\rangle \left\langle x\right|^X \otimes \rho_x^Q. \tag{9.182}$$

This only goes through because the system X is classical - so conditioning on it is sensible. This argument does not go through in the case of entangled inputs, i.e., for a general quantum-quantum channel.

9.2 General Quantum Channels

For a general quantum channel given by the CPTP map $\mathcal{N}$, the capacity is in fact

$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} \chi(\mathcal{N}^{\otimes n}), \tag{9.183}$$

where

$$\chi(\mathcal{E}) = \max_{\{p(x),\sigma_x\}} I(X : Q)_\omega, \tag{9.184}$$

for $\omega = \sum_x p(x) \left|x\right\rangle \left\langle x\right|^X \otimes \mathcal{N}(\sigma_x)^Q$. The limit in the formula reflects the fact that in general the quantum capacity is super additive, i.e., $\chi(\mathcal{N} \otimes \mathcal{M}) \geq \chi(\mathcal{N}) + \chi(\mathcal{M})$ always, and there are cases where this inequality is strict. This makes the formula quite difficult to work with, a fact also reflected by the optimisation problem contained in it being NP-hard to solve in the worst case. There are, however, several natural cases in which the quantum capacity is additive, meaning that $\chi(\mathcal{N}^{\otimes n}) = n\chi(\mathcal{N})$, and in these cases the formula becomes much easier to work with. Let us state some examples.

1. For entanglement-breaking channels, also called QCQ channels, or measure-and-prepare channels, the capacity is additive. These are channels of the form

$$N(\rho) = \sum_k Tr(\rho M_k)\sigma_k \tag{9.185}$$

for some measurement $\{M_k\}$ and states σ_k. Note that these are more general than CQ channels, for which the first measurement is in the computational basis.

2. Depolarising channel
$$\mathcal{N}(\rho) = (1-p)\rho + p\frac{I}{d}. \tag{9.186}$$
3. Erasure channel
$$\mathcal{N}(\rho) = (1-p)\rho + p\left|e\right\rangle\left\langle e\right|, \tag{9.187}$$
where $|e\rangle$ is some erasure symbol.
4. Unital qubit channels (such channels satisfy $\mathcal{N}(I) = I$ and have one qubit as their input and output).
5. Pure Loss Bosonic Channels. A Bosonic channel can be thought of as acting on the Hilbert space of a harmonic oscillator.

9.3 Random Access Coding

A natural application of these ideas comes in random access coding, specifically the related quantum no-go theorem. The task is as follows. Suppose Alice wishes to encode m bits $x^m \in \{0,1\}^m$ in an n-qubit quantum state ρ_x. She sends this to Bob. Bob wishes to retrieve just one of the bits, of his choosing, say i, by performing a measurement. He learns some bit $\hat{x}_i$ by performing a measurement, and the hope is that $\hat{x}_i = x_i$. One naturally asks to what extent this can be done reliably with various values of n and m.

At finite lengths, you can do a little better with quantum states than you can do classically. For example, clearly, if you encode 2 bits into 1 bit, there is only a 50% probability that Bob can learn a bit of his choosing. However, suppose that Alice encodes 00 into $|0\rangle$, 01 into $|+\rangle$, 10 into $|-\rangle$ and 11 into $|1\rangle$. Then, if Bob wishes to learn the first bit, he needs to distinguish
$$\frac{|0\rangle\langle 0| + |+\rangle\langle +|}{2} \text{ from } \frac{|-\rangle\langle -| + |1\rangle\langle 1|}{2} \tag{9.188}$$
and if he wishes to learn the second bit, he needs to distinguish
$$\frac{|0\rangle\langle 0| + |-\rangle\langle -|}{2} \text{ from } \frac{|+\rangle\langle +| + |1\rangle\langle 1|}{2}, \tag{9.189}$$
and in both cases he can succeed with probability $\cos^2 \pi/8 > 0.5$. Asymptotically, however, it turns out that he can essentially do just as well.

9.3.1 Nayak's No-Go Theorem

Theorem 9.3.1. *If any bit can be retrieved with probability $\geq 1-\epsilon$, then $n \geq m(1-H_2(\epsilon))$.*

To prove this, the following will be very useful.

Lemma 9.3.1. *Given states σ_0, σ_1 and a measurement $\{M_0, M_1\}$ which is good at distinguishing them, i.e., $Tr(M_b\sigma_b) \geq 1-\epsilon$ for each b, then*
$$S(\sigma) \geq \frac{S(\sigma_0) + S(\sigma_1)}{2} + 1 - H_2(\epsilon), \tag{9.190}$$
where
$$\sigma = \frac{\sigma_0 + \sigma_1}{2}. \tag{9.191}$$

Proof. To prove the lemma, let us define the CQ state

$$\rho_{XQ} = \frac{|0\rangle\langle 0| \otimes \sigma_0 + |1\rangle\langle 1| \otimes \sigma_1}{2} \tag{9.192}$$

for which we know that

$$I(X:Q) = S(\sigma) - \left(\frac{S(\sigma_0) + S(\sigma_1)}{2}\right). \tag{9.193}$$

Considering the CQ channel/measure scenario

$$X - Q - B \tag{9.194}$$

given by

$$x \mapsto \sigma_x \mapsto b, \tag{9.195}$$

we know from the converse of the HSW theorem that the mutual information between X and B is at most $I(X:Q)_\rho$:

$$I(X:B) \leq I(X:Q)_\rho. \tag{9.196}$$

However, since $I(X:B)$ agree with probability at least $1-\epsilon$, we have $I(X:B) \geq 1 - H_2(\epsilon)$, from which the result follows. □

Finally, we can prove Nayak's No-Go Theorem.

Proof. We define the CQ state

$$\rho = \frac{1}{2^m} \sum_{x \in \{0,1\}^m} |x\rangle\langle x|^X \otimes \rho_x^Q. \tag{9.197}$$

Then, we know that there is a measurement that is good at distinguishing the cases of x_{k+1} being 0 or 1. The lemma then tells us that

$$S(Q|X_1 \ldots X_k) \geq S(Q|X_1 \ldots X_{k+1}) + 1 - H_2(\epsilon), \tag{9.198}$$

because $S(Q|X_1 \ldots X_k)$ is the entropy of the state averaged over the values of X_{k+1}, and $S(Q|X_1 \ldots X_{k+1})$ is the average of the entropies over the values of X_{k+1}. Iterating this gives

$$S(Q) \geq m(1 - H_2(\epsilon)), \tag{9.199}$$

and we conclude via the observation that $n \geq S(Q)$, since Q is an n-qubit register. □

8.372 Quantum Information Science III **Fall 2024**

Lecture 11: October 10, 2024

Scribe: Louis Marquis *Quantum Sensing and Fisher Information*

11.1 Quantum Sensing

We have previously applied Holevo Information and relative entropy to classical information theory problems like hypothesis testing, channel encoding, and state tomography. We will now apply it to quantum sensing, which is a variant of state tomography.

Suppose that we have a magnetic field with unknown magnitude B and an electron (comprising a qubit), which results in the following Hamiltonian.

$$H = \frac{B}{2}Z \tag{11.200}$$

If there are N such particles, the total Hamiltonian is the sum of the individuals.

$$H = \sum_{i=1}^{N} \frac{B}{2}Z_i \tag{11.201}$$

The goal is to estimate B as precisely as possible. We wish to determine the a informationally-theoretic limit on such a precision.
A first attempt on the one-qubit scenario may involve evolving the state $|+\rangle$ with the Hamiltonian H, then measuring in the $\{|+\rangle, |-\rangle\}$ basis after time t. Denote $\phi = Bt$.

$$e^{-iHt}|+\rangle = \frac{1}{\sqrt{2}}(e^{-\frac{i\phi}{2}}|0\rangle + e^{\frac{i\phi}{2}}|1\rangle) \tag{11.202}$$

$$|\langle +| e^{-iHt}|+\rangle|^2 = |\langle +| \frac{1}{\sqrt{2}}(e^{-\frac{i\phi}{2}}|0\rangle + e^{\frac{i\phi}{2}}|1\rangle)|^2 = \cos^2\frac{\phi}{2} \tag{11.203}$$

Let random variable X represent the sign of the measured state. That is, $X = 1$ if $|+\rangle$ is measured, and $X = -1$ if $|-\rangle$ is measured. Clearly, the expected value is $E(X) = \cos^2\frac{\phi}{2} - \sin^2\frac{\phi}{2} = \cos\phi$. If the experiment is repeated many (N) times, the average X should approach the expected value $\bar{X}$, allowing an estimate ϕ (and in turn B).

$$\bar{x} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{11.204}$$

$$\phi = \arccos\bar{x} \tag{11.205}$$

$$B = \frac{\phi}{t} \tag{11.206}$$

We want to estimate the uncertainty of B. We can calculate the uncertainty Δx, which we define as the standard deviation of x. Then the uncertainty can be propogated over to get ΔB.

$$\Delta x = \sigma(X) = \sqrt{E(X^2) - E(X)^2} = \sqrt{1 - \cos^2\phi} = |\sin(\phi)| \tag{11.207}$$

$$\Delta\bar{x} = \sigma(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{\sigma(X)}{\sqrt{N}} = \frac{|\sin(\phi)|}{\sqrt{N}} \tag{11.208}$$

$$\Delta\phi = \frac{\Delta\bar{x}}{|\frac{d\bar{x}}{d\phi}|} = \frac{\frac{|\sin(\phi)|}{\sqrt{N}}}{|\sin\phi|} = \frac{1}{\sqrt{N}} \tag{11.209}$$

$$\Delta B = \frac{\Delta\phi}{t} = \frac{1}{\sqrt{N}t} \tag{11.210}$$

Particularly remarkable about this result is that neither $\Delta\phi$ nor ΔB depend on anything besides B and t, including anything relating to phase. This means that the experiment is equally precise regardless of the phase of the starting state.
$\frac{1}{\sqrt{N}}$ is referred to as the standard quantum limit (SQL), or the shot-noise limit.
This first attempt consisted of independent measurements of N different qubits. A better precision can be achieved by entangling the qubits first into the cat state. In practice, the cat state is more vulnerable to noise but we can ignore this aspect for now. We evolve this state with the N qubit Hamiltonian.

$$e^{-iHt}\frac{1}{\sqrt{2}}(|0\rangle^{\otimes N} + |1\rangle^{\otimes N}) = \frac{1}{\sqrt{2}}(e^{-\frac{i\phi N}{2}}|0\rangle^{\otimes N} + e^{\frac{i\phi N}{2}}|1\rangle^{\otimes N}) \tag{11.211}$$

We notice that the phase shift increases by a factor of N (relative to the first attempt), but this time we only run the multi-qubit experiment once (instead of the single-qubit one N times). This allows a quick calculation of the new precision.

$$\Delta B = \frac{1}{Nt} \tag{11.212}$$

This is referred to as the Heisenberg limit, as it parallels the Heisenberg Uncertainty Principle.

11.2 Hamiltonian Learning

The problem of quantum sensing is a specific case of the more general problem of Hamiltonian learning, in which the Hamiltonian. In theory, the maximally general Hamiltonian can have 4^N unknown parameters, but such a problem isn't very useful nor interesting. In Hamiltonian learning, we know the general structure of the Hamiltonian as the linear combination of a set of terms $\{h_i\}$ and seek the specific parameters $\{b_i \in \mathbb{R}\}$ of this combination.

$$H = \sum_i \beta_i h_i \tag{11.213}$$

In this problem, one must prepare a state, evolve it with the Hamiltonian, and measure it to gain information on $\{h_i\}$.

Alternatively, one might be given the Gibbs state for a Hamiltonian $\frac{e^{-\frac{H}{T}}}{\mathrm{tr}\left(e^{-\frac{H}{T}}\right)}$ and must determine $\{h_i\}$ by measuring the state. This problem is also referred to as Hamiltonian learning.
However, Hamiltonian learning is a very complicated topic since there are so many strategies that can be considered. So for the rest of this lecture, we will focus on a simpler problem called parameter estimation from states. This problem is simple enough to have a full solution, and this solution reveals insights on Hamiltonian learning.

11.3 Parameter Estimation from States

In this problem, there is an unknown parameter θ that determines the distribution $p_\theta(x)$ of observation x. The goal is to output an optimal estimate $\hat{\theta}$ after observing x. It is assumed that $p_\theta(x)$ is continuous and differentiable over θ.
A simpler version of the problem involves distinguishing $p_\theta(x)$ from $p_0(x)$ for θ close to 0. In this case, we can simply use a likelihood ratio test for hypothesis testing. Define W_n as the logarithm of such a ratio.

$$W_n(x^n) = \log \frac{\prod_{i=1}^n p_\theta(x_i)}{\prod_{i=1}^n p_0(x_i)} \tag{11.214}$$

We have seen before a bound on the expectation of W_n for the $x^n \leftarrow p_0^n$ case.

$$E_{x^n \leftarrow p_0^n}(W_n) \leq 0 \tag{11.215}$$

We can also calculate the $x^n \leftarrow p_\theta^n$ case.

$$E_{x^n \leftarrow p_\theta^n}(W_n) = \sum_{x^n} p_\theta^n(x^n) \log \frac{\prod_{i=1}^n p_\theta(x_i)}{\prod_{i=1}^n p_0(x_i)} \tag{11.216}$$

$$= \sum_{i=1}^n \sum_{x^n} p_\theta^n(x^n) \log \frac{p_\theta(x_i)}{p_0(x_i)} \tag{11.217}$$

$$= \sum_{i=1}^n \sum_{x_i} p_\theta(x_i) \log \frac{p_\theta(x_i)}{p_0(x_i)} \tag{11.218}$$

$$= nD(p_\theta||p_0) \tag{11.219}$$

We can estimate the relative entropy $D(p_\theta||p_0)$ for small θ by expanding it as a power series. At $\theta = 0$, we know that it is zero and symmetric, so the constant and linear terms must be zero. Therefore, the first potentially nonzero term is the quadratic term, which we denote F.

$$E(W_n) = nD(p_\theta||p_0) = n(0 * 1 + 0 * \theta + F * \frac{\theta^2}{2} + O(\theta^3)) = \frac{nF\theta^2}{2} + O(\theta^3) \tag{11.220}$$

We can restate this a direct formula for F, which we call the Fisher information.

$$F = \partial_\theta^2 D(p_\theta||p_0)|_{\theta=0} \tag{11.221}$$

The Fisher information has several equivalent forms, which are useful but will not be derived in this lecture.

$$F = \sum_x p_\theta(x)(\partial_\theta \log p_\theta(x))^2|_{\theta=0} \tag{11.222}$$

$$= \sum_x \frac{(\partial_\theta p_\theta(x))^2}{p_\theta}|_{\theta=0} \tag{11.223}$$

We now show that the Fisher information indicates whether p_0^n and p_θ^n can be reliably distinguished. We define this condition as the expectation of W_n under p_θ^n being greater than its standard deviation under either p_θ^n or p_0^n. Such a condition can be found by calculating the variance with a factor of

$\frac{1}{n}$ for convenience.

$$\frac{1}{n}\sigma^2_{x^n \leftarrow p_\theta^n}(W_n) = \sum_x p_\theta(x)(\log \frac{p_\theta(x)}{p_0(x)})^2 - D(p_\theta, p_0)^2 \tag{11.224}$$

$$= \sum_x p_\theta(x)(\theta \partial_\theta \log p_\theta + O(\theta^2))^2 - O(\theta^4) \tag{11.225}$$

$$= \theta^2 \sum_x p_\theta(x)(\partial_\theta \log p_\theta)^2 + O(\theta^3) \tag{11.226}$$

$$= F\theta^2 + O(\theta^3) \tag{11.227}$$

To reliably distinguish the two distributions, the expectation must greater than the standard deviation. We ignore $O(\theta^3)$ elements.

$$\frac{nF\theta^2}{2} = E(W_n) \geq \sqrt{\sigma^2(W_n)} = \sqrt{nF\theta^2} \implies \theta \geq \frac{1}{\sqrt{nF}} \tag{11.228}$$

This is the minimum θ at which one can reliably distinguish p_θ, p_0.

11.4 Cramer-Rao Bound

We have found the minimum θ whose distribution can be reliably distinguished from that of 0. This quantity also happends to also be the Cramer-Rao Bound.

Theorem 11.4.1. *Define an estimator $\hat{\theta}(x^n)$ as locally unbiased if the expectation of the estimate $\hat{\theta}$ is approximately θ when θ is close to θ_0. θ_0 will almost always be set to 0 in practice.*

$$E_{x^n \leftarrow p_\theta^n}(\hat{\theta}) = \theta + O((\theta - \theta_0)^2) \tag{11.229}$$

If $\hat{\theta}(x^n)$ is locally unbiased, then $\sigma(\hat{\theta}) \geq \frac{1}{\sqrt{nF}}$.

Proof. We wish to find a lower bound for the variance of $\hat{\theta}$. We can begin by noticing that $E(\hat{\theta}) = \theta + O(\theta^2)$ and since θ is small, this term can be ignored in the variance of $\hat{\theta}$.

$$\sigma^2(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 \approx E(\hat{\theta}^2) = \sum_{x^n} p_\theta^n(x^n)\hat{\theta}(x^n)^2 \tag{11.230}$$

We can then take a derivative of the locally unbiased condition.

$$1 = \partial_\theta E(\hat{\theta})|_{\theta=0} \tag{11.231}$$

$$= \partial_\theta \sum_{x^n} p_\theta^n(x^n)\hat{\theta}(x^n)|_{\theta=0} \tag{11.232}$$

$$= \sum_{x^n} \partial_\theta p_\theta^n(x^n)|_{\theta=0}\hat{\theta}(x^n) \tag{11.233}$$

$$= E_{x^n \leftarrow p_\theta^n(x^n)}(\frac{\partial_\theta p_\theta^n(x^n)|_{\theta=0}}{p_\theta^n(x^n)}\hat{\theta}(x^n)) \tag{11.234}$$

$$= E_{x^n \leftarrow p_\theta^n(x^n)}(\partial_\theta \log p_\theta^n(x^n)|_{\theta=0}\hat{\theta}(x^n)) \tag{11.235}$$

One can define an inner product $a \cdot b$ over functions a, b of x^n and apply the Cauchy-Schwarz Inequality over these functions.

$$a \cdot b := E_{x^n \leftarrow p_\theta(x^n)} a(x^n)b(x^n) \tag{11.236}$$

$$(a \cdot b)^2 \leq (a \cdot a)(b \cdot b) \tag{11.237}$$

In our current case, $a(x^n) = \partial_\theta \log p_\theta^n(x^n)|_{\theta=0}$ and $b(x^n) = \hat{\theta}(x^n)$.

$$\begin{aligned}
1^2 &= E_{x^n \leftarrow p_\theta^n(x^n)}(\partial_\theta \log p_\theta^n(x^n)|_{\theta=0}\hat{\theta}(x^n))^2 && (11.238)\\
&= E_{x^n \leftarrow p_\theta^n(x^n)}(a(x^n)b(x^n))^2 && (11.239)\\
&= (a \cdot b)^2 && (11.240)\\
&\leq (a \cdot a)(b \cdot b) && (11.241)\\
&= E_{x^n \leftarrow p_\theta^n(x^n)}a(x^n)E_{x^n \leftarrow p_\theta^n(x^n)}b(x^n) && (11.242)\\
&= E_{x^n \leftarrow p_\theta^n(x^n)}(\partial_\theta \log p_\theta^n(x^n)|_{\theta=0})^2 E_{x^n \leftarrow p_\theta^n(x^n)}(\hat{\theta}(x^n))^2 && (11.243)\\
&= nF\sigma^2(\hat{\theta}) \implies && (11.244)\\
\sigma^2(\hat{\theta}) &\geq \frac{1}{nF} && (11.245)
\end{aligned}$$

□

11.5 Quantum Fisher Information

We will now define the quantum version of Fisher information. In the quantum version of the parameter estimation problem, the parameter θ determines a density matrix ρ_θ from which to sample x, instead of a probability distribution. It is given that $\rho_\theta > 0$ when θ is near $\theta_0 \approx 0$.
We will also now define some new super-operators on matrices.

$$\text{Mult}_\rho(X) = \frac{1}{2}(\rho X + X\rho) \tag{11.246}$$

$$\text{Div}_\rho = \text{Mult}_\rho^{-1} \tag{11.247}$$

$$L_{\rho,\theta} = \text{Div}_{\rho_\theta}(\partial_\theta \rho_\theta) \tag{11.248}$$

$$\partial_\theta \rho_\theta = \text{Mult}_{\rho_\theta}(L_{\rho,\theta}) \tag{11.249}$$

With these super-operators, we can define the quantum Fisher information.

$$F_Q = \text{tr}\big(\rho L^2\big) \tag{11.250}$$

Analogously to the classical Fisher information, the quantum Fisher information relates to the second derivative of the quantum relative entropy.

$$F_Q = \partial_\theta^2 D(\rho_\theta || \rho_0)|_{\theta=0} \tag{11.251}$$

We can also relate the quantum Fisher information to the classical Fisher information. To do this, define a set of measurement operators $\{M_x\}_{x \in X}$ with $\sum_{x \in X} M_x = I$. We can calculate the Fisher information of the distribution of the measurement, which is $\rho_\theta(x) = \text{tr}(\rho_\theta M_x)$.

$$\begin{aligned}
F_M &= F(\rho_\theta(x)) && (11.252)\\
&= \sum_{x \in X} \text{tr}(\rho_\theta M_x)\left(\frac{\partial_\theta \,\text{tr}(\rho_\theta M_x)}{\text{tr}(\rho_\theta M_x)}\right)^2 && (11.253)\\
&= \sum_{x \in X} \text{tr}(\rho_\theta M_x)\left(\frac{\text{Re}\{\text{tr}(\rho_\theta L M_x)\}}{\text{tr}(\rho_\theta M_x)}\right)^2 && (11.254)\\
&\leq \sum_{x \in X} \text{tr}(\rho_\theta M_x)\left(\frac{|\,\text{tr}(\rho_\theta L M_x)|}{\text{tr}(\rho_\theta M_x)}\right)^2 && (11.255)
\end{aligned}$$

We used the fact that $\partial_\theta \rho_\theta$ is essentially the Hermitian part of $\rho_\theta L_{\rho,\theta}$, so $\partial_\theta \operatorname{tr}(\rho_\theta M_x) = \operatorname{Re}\{\operatorname{tr}(\rho_\theta L M_x)\}$. We can now use the quantum Cauchy-Schwarz Inequality on $|\operatorname{tr}(\rho_\theta L M_x)|$. Specifically, the inequality states that $|\operatorname{tr}(AB)| \leq \sqrt{\operatorname{tr}(A^\dagger A)\operatorname{tr}(B^\dagger B)}$.

$$|\operatorname{tr}(\rho_\theta L M_x)| = |\operatorname{tr}\Big(\sqrt{\rho_\theta} L \sqrt{M_x}\sqrt{M_x}\sqrt{\rho_\theta}\Big)| \tag{11.256}$$

$$\leq \sqrt{\operatorname{tr}(\rho_\theta L M_x L)\operatorname{tr}(\rho_\theta M_x)} \tag{11.257}$$

We can finally substitute this inequality back into F_M.

$$F_M \leq \sum_{x\in X} \operatorname{tr}(\rho_\theta M_x)\Big(\frac{|\operatorname{tr}(\rho_\theta L M_x)|}{\operatorname{tr}(\rho_\theta M_x)}\Big)^2 \tag{11.258}$$

$$\leq \sum_{x\in X} \operatorname{tr}(\rho_\theta L M_x L) \tag{11.259}$$

$$= \operatorname{tr}\big(\rho_\theta L^2\big) \tag{11.260}$$

$$= F_Q \tag{11.261}$$

In summary, the quantum Fisher information of a density matrix is at least the classical Fisher information of distribution of measurement outcomes on that density matrix. The inequality is tight if the measurement is done in the eigenbasis of L.