Ph/CS 219c
2 March 2011
Accessible information

Note Title 3/2/2011

Mixing and measurement:

Recall that Shannon entropy is Schur concave — If the probability vector (p_x) is majorized by (q_y) (i.e. $(p_x) \prec (q_y)$), then $H(X) \geq H(Y)$. Both relations reflect that "X is more random than Y." Thus a doubly stochastic channel

$$p_x = \sum_y D_{xy} q_y \qquad \text{(where rows and columns of } D \text{ sum to 1)}$$

does not decrease entropy.

Von Neumann entropy $H(\rho)$ of density operator ρ is Shannon entropy of $\lambda(\rho)$ — the vector of eigenvalues of ρ. Recall that we have shown (as a consequence of the HJW thm.) that any ρ is realized as an ensemble of pure states

$$\{|\psi_x\rangle, p_x\} \longrightarrow \rho = \sum_x p_x |\psi_x\rangle\langle\psi_x|$$

Then $p \prec \lambda(\rho)$; therefore $H(\rho) \leq H(X)$. The VN entropy of density operator is no larger than the Shannon entropy of the mixture. The entropies are equal iff the states $\{|\psi_x\rangle\}$ are mutually orthogonal. In general

$$\sqrt{p_x}\,|\psi_x\rangle = \sum_a V_{xa} \sqrt{\lambda_a}\,|a\rangle$$

(V: unitary; $|a\rangle$: eigenvectors of ρ)

$$\Rightarrow \quad p_x = \sum_a |V_{xa}|^2 \lambda_a$$

($|V_{xa}|^2$: doubly stochastic D_{xi})

As we'll soon see, when we perform a measurement on the system with density operator ρ, the amount of information we can gain about ρ

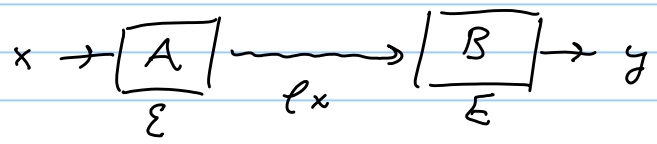
is no more than $H(\rho)$ bits. So mixing of nonorthogonal states is irreversible: that is, information about which state was prepared is irretrievably lost.

If we perform the orthogonal measurement that projects onto the basis $\{|y\rangle\}$, then outcome y occurs with probability

$$q_y = \langle y|\rho|y\rangle = \sum_a \lambda_a |\langle a|y\rangle|^2$$

where $\rho = \sum_a \lambda_a |a\rangle\langle a|$; since $|\langle a|y\rangle|^2$ is doubly stochastic, $(q_y) \prec \lambda(\rho)$ and hence $H(Y) \geq H(\rho)$, with equality iff the measurement is in basis in which ρ is diagonal. Any other orthogonal measurement produces an outcome that is "more random" than the measurement in the diagonal basis.

Let's ask a sharper question — how much information can we gain by making a measurement? Consider a game



Alice prepares a state by sampling from the ensemble $\mathcal{E} = \{\rho_x, p(x)\}$. Bob performs POVM with Kraus operators $E = \{E_y\}$, $\sum E_y^\dagger E_y = I$. Then conditional prob of outcome y if state prepared is ρ_x is

$$p(y|x) = \mathrm{tr}(E_y^\dagger E_y \rho_x)$$

and joint distribution is $p(x,y) = p(y|x)p(x)$

Averaged over Alice's prep. and Bob's outcome, Bob's information gain about Alice's state

is $I(X;Y)$. Bob's best strategy (his optimal measurement) maximizes his information gain. This maximum value of the mutual information is a property of the ensemble, which we call the "accessible information" of the ensemble:

$$Acc(\mathcal{E}) = \max_{\mathcal{E}} I(X;Y).$$

If the states $\{\rho_x\}$ are mutually orthogonal, we may choose $\{E_y\}$ to be projectors onto supports of these states; then

$$p(x|y) = \delta_{x,y} \Rightarrow H(X|Y) = 0 \text{ and } I(X;Y) = H(X).$$
$$= \langle -\log p(x|y) \rangle$$

In this case, the optimal measurement determines Alice's preparation perfectly. But if ensemble is not orthogonal, then

$$H(X|Y) > 0 \quad \text{and} \quad I(X;Y) < H(X);$$

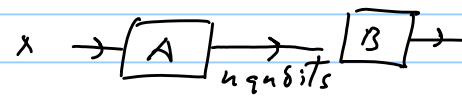
some info about what Alice prepared is inaccessible.

There is no useful general formula for $Acc(\mathcal{E})$, but we can derive a useful upper bound:

For $\mathcal{E} = \{|\psi_x\rangle, p(x)\}$ (an ensemble of pure states),

$$Acc(\mathcal{E}) \le H(\rho), \text{ where } \rho = \sum p(x) |\psi_x\rangle\langle\psi_x|$$

This (a special case of) the "Holevo bound", and it is not tight in general; it can be generalized also to an alphabet of mixed states, as we'll discuss later. Note that if Alice tries to encode a message x by choosing a state $|\psi_x\rangle$ of n qubits,

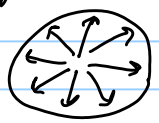


then

$$I(X;Y) \le H(\rho^{(n)}) \le n,$$

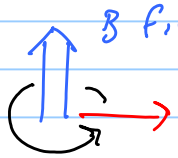
since $H(\rho) \le \log_2 d$ and here $d = 2^n$

Thus, a qubit can convey at most one bit of classical info from Alice to Bob. Alice might try to encode more classical info by



choosing from a large alphabet of pure states — i.e. vectors on the Bloch sphere. But these states are imperfectly distinguishable -- so Bob cannot extract more than a bit about what Alice prepared.

Why should we care about how well information can be encoded in nonorthogonal states? The players in the game might be, rather than "Alice" and "Bob," "Nature" and "experimenter."



For example, experimenter wants to measure magnetic field B, so he prepares spin in a known initial state, then in a specified time the spin precesses by an a priori unknown amount. The experimenter then wants to collect info about the direction in which the spin is pointing — i.e. he wants to distinguish nonorthogonal states. If the spins decohere, then he will need to distinguish among an alphabet of mixed states.

To derive the Holevo bound (and other interesting consequences) we exploit properties of Von Neumann entropy. In some ways VN entropy is like Shannon entropy, but in some important ways it is different.

Subadditivity: $H(AB) \leq H(A) + H(B)$

(a HW exercise)

which implies

$$I(A;B) = H(A) + H(B) - H(AB) \geq 0 ;$$

quantum mutual information is nonnegative.

$I(A;B)$ expresses how strongly correlated are the systems A and B. It vanishes iff the state of AB is a product state

$$\rho_{AB} = \rho_A \otimes \rho_B = \sum_{a,b} |a,b\rangle \lambda_a \chi_b \langle a,b|$$

i.e. the vector of eigenvalues is a product distribution

$$\Rightarrow \quad H(AB) = H(A) + H(B)$$

An interesting difference between Shannon and VN entropy concerns conditional entropy. Classically

$$H(XY) \geq H(Y) \Rightarrow$$

$$H(X|Y) = H(XY) - H(Y) \geq 0$$

Recall that $H(X|Y) = \langle -\log p(x|y) \rangle = \left\langle -\log \frac{p(xy)}{p(y)} \right\rangle$

quantifies our remaing uncertainty about X once Y is known.

But quantumly we can have $H(AB) < H(B)$

$$\Rightarrow H(A|B) = H(AB) - H(B) < 0$$

For example, if AB is pure, then $H(AB) = 0$ and $H(B) = H(A) = E$ is entanglement of A and B

$$\Rightarrow H(A|B) = -E < 0$$

It is as though our remaining uncertainty about A when B is known is negative — we are "more than certain." Actually, negative uncertainty has a sensible operational interpretation, which we'll come to later.

Strong Subadditivity

Classically, Shannon mutual info satisfies "strong subadditivity":

$$I(X;YZ) \geq I(X;Y)$$

"Obviously" the info you gain about X when you know Y and Z is no less than when you know only Y! Strong subadditivity follows from the "chain rule" for mutual info (which holds classically or quantumly). Note that

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

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

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

which implies the identity

$$I(X;YZ) = I(X;Y) + I(X;Z|Y) \qquad \text{("chain rule")}$$

But classically it is also obvious that

$$I(X;Z|Y) = \sum_y p(y)\, I(X;Z|y) \geq 0$$

(a convex comb. of nonnegative quantities)

From which follows: $I(X;YZ) \geq I(X;Y)$

Quantumly it is also true that $I(A;C|B) \geq 0$,

and therefore $\boxed{I(A;BC) \geq I(A;B)}$

But in the quantum case strong subadditivity is a rather deep theorem — there is no known elementary proof. (Later, we will give an operational proof, based on state merging, using "decoupling" and the theory of typical subspaces.)

There are a few other ways to express strong subadditivity that are sometimes useful:

$$H(A) - H(A|BC) \geq H(A) - H(A|B) \implies$$

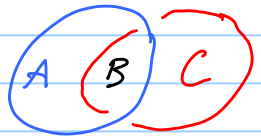
$$H(A|B) \geq H(A|BC)$$

(A becomes less uncertain when we know C as well as B)

Also:

$$H(A) + H(B) - H(AB) \leq H(A) + H(BC) - H(ABC)$$

$$\Rightarrow \quad H(ABC) + H(B) \leq H(AB) + H(BC)$$



If AB and BC are two overlapping systems, then ABC is their union, and B is their intersection. (This reduces to subadditivity when intersection B is trivial.)

Monotonicity

One important consequence of strong subadditivity (SSA) is monotonicity of quantum mutual information. This means that an operation applied to B cannot increase the mutual information of A and B.

To derive monotonicity, we recall that a TPCP map $B \to B'$ can be realized by an isometry $B \to B'E$, where E is a suitable "environment." Furthermore, the VN entropy is invariant under a unitary change of basis: $H(\rho) = H(U\rho U^\dagger)$. Therefore

$$H(AB) = H(AB'E) \text{ and } H(B) = H(B'E) \Rightarrow$$

$$I(A;B)_{\text{Before}} = I(A;B'E)_{\text{After}}$$

Then from SSA we have $I(A;B'E) \geq I(A;B')$

$$\Rightarrow \quad \boxed{I(A;B)_{\text{Before}} \geq I(A;B')_{\text{After}}}$$

This makes sense: an operation on B cannot increase the correlation of B with A (though it can reduce the correlation).

Holevo Bound

One important consequence of SSA is the Holevo bound. In the accessible information game, we consider a three part system. Q is the quantum system that Alice prepares and Bob measures. Alice records the state that she prepares in register X and Bob records his measurement outcome in the register Y. We are interested in the mutual info $I(X;Y)$ of Alice's record and Bob's record. The joint state of XQ after Alice prepares is

$$\rho_{XQ} = \sum_x p(x) |x\rangle\langle x| \otimes \rho_x ,$$

which becomes after Bob measures:

$$\rho_{XQ'Y'} = \sum_{x,y} p(x) |x\rangle\langle x| \otimes E_y \rho_x E_y^\dagger \otimes |y\rangle\langle y|$$

$$\Rightarrow \rho_{XY'} = \sum_{x,y} p(x,y) |x\rangle\langle x| \otimes |y\rangle\langle y| \quad \text{where } p(x,y) = p(x)p(y|x).$$

SSA says that, after the measurement

$$I(X;Y')_{after} \leq I(X;Q'Y')_{after}$$

Furthermore, since the measurement is an operation applied to QY, monotonicity of mutual information implies

$$\boxed{I(X;Q'Y')_{after} \leq I(X;Q)_{before} \equiv \chi(\mathcal{E})}$$

where $\chi(\mathcal{E})$ is a property of the ensemble.

So now we should compute $I(X;Q) = H(Q) - H(Q|X)$:

$H(Q) = H(\rho)$ where $\rho = \sum_x p(x)\rho_x$ – the VN entropy of the density operator for ensemble.

$$H(Q|X) = \sum_x p(x) H(\rho_x)$$

– the entropy of the signal state conditioned on the preparation, averaged over the ensemble. To compute it explicitly:

$H(Q|X) = H(XQ) - H(X)$, where $H(X)$ is Shannon entropy of the ensemble, and

$$\rho_{XQ} = \sum_x p(x) (|x\rangle\langle x| \otimes \rho_x)$$

$$\Rightarrow H(XQ) = -\sum_x \mathrm{tr}[p(x)\rho_x \log p(x)\rho_x]$$
$$= H(X) - \sum_x p(x) \mathrm{tr}\,\rho_x \log \rho_x = H(X) + \sum_x p(x) H(\rho_x)$$

Therefore, $H(Q|X) = \sum_x p(x) H(\rho_x)$ and

$$I(X';Y') \le I(X,Q) = H(Q) - H(Q|X)$$

$$= H(\rho) - \langle H(\rho_x)\rangle$$

$H(\rho)$: VN entropy of ensemble

$\langle H(\rho_x)\rangle$: average VN entropy of the alphabet

The quantity $\chi(\mathcal{E}) = H(\rho) - \langle H(\rho_x)\rangle$

is called the "Holevo Chi" (or Holevo information) of the ensemble $\mathcal{E} = \{p(x), \rho_x\}$ and we have derived $Acc(\mathcal{E}) = \max_{\mathcal{E}} I(X;Y) \le \chi(\mathcal{E})$

For an ensemble of pure states, $H(\rho_x) = 0$ for all x, and

$$Acc(\mathcal{E}) \le H(\rho)$$

To summarize the argument succinctly,

$$I(X;Y)_{after} \overset{SSA}{\le} I(X;QY)_{after} \overset{Mono}{\le} I(X;Q)_{before} = \chi(\mathcal{E}),$$

where "after" refers to after Bob's measurement and before refers to before Bob's measurement. The first inequality follows from strong subadditivity, and the second from monotonicity of mutual information (itself a consequence of SSA).

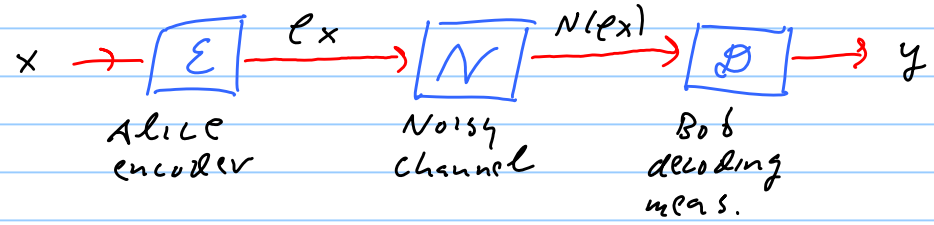
Since Holevo Chi is itself an instance of quantum mutual information, we can apply monotonicity to χ. If a channel $\mathcal{N}$ maps Q to Q', then

$$\chi' = I(X;Q') \leq I(X,Q) = \chi;$$

the channel cannot increase Holevo Chi of an ensemble

Classical Capacity of a Quantum Channel

Accessible information, and hence also Holevo Chi, are relevant for sending classical information through a noisy quantum channel



Alice encodes classical message by preparing quantum state ρ_x and Bob receives noisy version $\mathcal{N}(\rho_x)$ of the state. Bob measures and tries to infer x. For prob distribution $p(x)$ on the messages, Bob's optimal measurement achieves the accessible information of the ensemble

$$\mathcal{E} = \{\mathcal{N}(\rho_x), p(x)\}$$

which is bounded above by Holevo Chi of the ensemble

$$\chi(\mathcal{E}) = I(X;B)$$

– the quantum mutual information of the state

$$\rho_{XB} = \sum_x p(x)\, |x\rangle\langle x| \otimes \mathcal{N}(\rho_x)$$

In a single shot, the optimal meas. does not achieve Holevo chi, in general.

But if we use the channel many times, there is a capacity theorem saying that we can achieve X bits per letter asymptotically.

Before discussing achievability, consider the converse of the coding theorem, the upper bound on capacity. It is useful to recall the classical case, where the channel is the conditional prob. distribution $p(y|x)$. Suppose we use the channel n times, where there are 2^{nR} codewords (hence R bits conveyed per use of the channel) Let $\tilde{X}^n$ denote the uniform distribution on these codewords, which induces joint distribution $(\tilde{X}^n, \tilde{Y}^n)$ on channel inputs and outputs. The mutual information this distribution is

$$I(\tilde{X}^n; \tilde{Y}^n) = H(\tilde{X}^n) - H(\tilde{X}^n|\tilde{Y}^n).$$

But $H(\tilde{X}^n) = nR$ for the uniform distribution on codewords, and $H(\tilde{X}^n|\tilde{Y}^n) \to 0$ as $n \to \infty$ if the output can be decoded with negligible error probability; thus

$$R = \frac{1}{n}\left[I(\tilde{X}^n; \tilde{Y}^n) + \varepsilon\right]$$

The achievable rate, then, is bounded above by

$$R = \max_{X^n} \frac{1}{n} I(X^n; Y^n)$$

(since the uniform distribution on codewords is a special case of a distribution X^n)

We can express

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n),$$

and the channel acts independently on each letter, so that

$$H(Y^n|X^n) = \sum_i \langle -\log p(y_i|x_i) \rangle = \sum_{i=1}^n H(Y_i|X_i)$$

Entropy is subadditive $\Longrightarrow$

$$H(Y^n) \le \sum_i H(Y_i) \quad \text{, and therefore}$$

$$\frac{1}{n} I(X^n; Y^n) \le \frac{1}{n}\left(\sum_i I(X_i; Y_i)\right) \le \max_X I(X; Y),$$

which implies

$$\lim_{n\to\infty} \max_{X^n} \frac{1}{n} I(X^n; Y^n) = \max_X I(X; Y) = C$$

The left-hand side is a valid expression for capacity (it is achievable), but not very useful, since it involves an arbitrary number of channel uses; we say it is a "regularized" expression. But because $I(X^n; Y^n)$ is subadditive, it reduces to a "single-letter formula" involving just one use of the channel

Now consider sending classical information over a quantum channel, where the quantum channel is used n times. For a code with rate R, consider the uniform distribution over the 2^{nR} codewords. If Bob can decode with negligible probability of error, then his optimal measurement can identify the codeword sent, and his information gain is $I(\tilde{X}^n; \tilde{Y}^n) = nR - \varepsilon$ (where $\varepsilon \to 0$ as $n \to \infty$)

By the Holevo bound, then, the rate satisfies

$$R = \frac{1}{n}\left[I(\tilde{X}^n; \tilde{Y}^n) + \varepsilon\right] \le \frac{1}{n}\left[\max_{X^n} I(X^n; B^n) + \varepsilon\right]$$

If we define the classical capacity C of the quantum channel as the maximum rate that can be achieved with error prob. $\to 0$ as $n \to \infty$, then

$$C(\mathcal{N}) \leq \max_{X^n} \left[\frac{1}{n} I(X^n; B^n) \right]$$

In fact, this expression really is achievable using random coding, so it is a regularized expression for the classical capacity. Can it be reduced to a single-letter formula?

$$I(X^n; B^n) = H(B^n) - H(B^n|X^n)$$

subadditivity of VN entropy implies

$$H(B^n) \leq \sum_{i=1}^{n} H(B_i),$$

where B_i is the marginal density operator for the ith letter that Bob receives. But what can we say about $H(B^n|X^n)$?

If Alice uses the channel n times, then in principle she could choose her quantum signals to be entangled states of the n letters she sends, and in that case the signals that Bob receives could also be entangled. But suppose that Alice decides to send codewords that are product states (though of course the signals sent in channel uses $i = 1, 2, \ldots, n$ could be classically correlated). Bob will still want to use the optimal POVM, which might act collectively on the n letters he receives, but $I(X^n; B^n)$ is still an upper bound on the information Bob can gain. In general, then, Bob receives states chosen from some ensemble of product states

$$\mathcal{E} = \left\{ \rho_{x_1} \otimes \rho_{x_2} \otimes \cdots \otimes \rho_{x_n},\ p(x_1, x_2, \ldots, x_n) \right\}$$

(where $p(x)$ may be correlated among $x_1, x_2, \ldots x_n$).

Then $$H(B^n|X^n) = \sum_{x_1 \cdots x_n} p(x_1, \ldots, x_n) H(\rho_1 \otimes \cdots \otimes \rho_n)$$

since the entropy of a product state is additive,

$$H(\rho_{x_1} \otimes \cdots \otimes \rho_{x_n}) = \sum_i H(\rho_{x_i}),$$ we have

$$H(B^n|X^n) = \sum_{x_1} p_1(x_1) H(\rho_{x_1}) + \cdots + \sum_{x_n} p_n(x_n) H(\rho_{x_n})$$

$$= \sum_{i=1}^{n} H(B_i|X_i)$$

(where $p_i(x_i)$ is the marginal prob. distribution for the ith letter. So for the special case of product codewords, the mutual information is subadditive

$$I(X^n; B^n)\Big|_{\substack{\text{product} \\ \text{states}}} \leq \sum_i I(X_i; B_i) \leq n \max_X I(X;B)$$

So our upper bound on the capacity (which is achievable with product state codewords) becomes a single-letter formula:

$$C_1 = \max_X I(X;B) \equiv \chi(\mathcal{N})$$

(which is a property of the channel $\mathcal{N}$). C_1 is the "product state classical capacity" of the channel $\mathcal{N}$.

To show that C_1 is really an achievable rate, we use random coding. For a given input ensemble $\{\rho_x, p(x)\}$ we generate n-letter codewords by sampling from the ensemble n times; when ρ_x is sent, Bob receives $\mathcal{N}(\rho_x)$. In Shannon's case, we could say that with high prob Bob receives one of at least $2^{n(H(Y)-\delta)}$ typical messages, and that

for each message sent by Alice, Bob receives one of $2^{n(H(Y|X)+\delta)}$ Typical messages. If Alice sends one of 2^{nR} messages, the prob of a decoding error, because the message received is in more than one decoding sphere is

$$\text{error prob} \leq 2^{nR} \frac{2^{nH(Y|X)+\delta}}{2^{n(H(Y)-\delta)}}$$

$$\leq 2^{n(R-I(X;Y)+2\delta)}$$

which approaches zero as $n \to \infty$ for $R < I - 2\delta$

Quantumly, Bob receives quantum message in a typical subspace of dimension at least $2^{n(H(B)-\delta)}$ and for each message sent by Alice, the signal is in a typical subspace of dimension $2^{n(H(B|X)+\delta)}$

To give an honest argument we should specify Bob's decoding POVM and estimate its error probability. But that is rather technical, so suffice it to say that the probability that signal received is accidentally in the decoding subspace of another signal is

$$\text{error prob} < 2^{nR} \frac{2^{n(H(B|X)+\delta)}}{2^{n(H(B)-\delta)}}$$

$$= 2^{n(R-I(X;B)+2\delta)}$$

so we can achieve the rate $R = I(X;B)$ asymptotically.

But is the capacity C really the same as the product state capacity C_1? It would be if we could show $I(X^n; B^n) \leq \sum_i I(X_i; B_i)$ in general, for entangled signals as well as product signals. However, a recent discovery is that the Holevo chi is not subadditive — there are channels such that $\chi(N_1 \otimes N_2) > \chi(N_1) + \chi(N_2)$. We say that $\chi(N)$ is "superadditive". As a result, our understanding of the classical capacities of quantum channels is far from complete. We have only a regularized formula, not a single-letter formula.

Quantum Channel Capacity

The formula $C = \max_X I(X;Y)$ for capacity of a classical channel is quite robust. For example, the capacity does not increase if we allow sender and receiver to share randomness, or if we allow feedback from the receiver to the sender. But quantumly the situation is more complicated. For example, shared entanglement boosts the quantum capacity, as does classical communication from receiver to sender. So there are a variety of different natural notions of quantum capacity, all with different capacity formulas.

Perhaps the most natural quantum capacity (unassisted by entanglement, and with one-way quantum communication) is this:

Alice encodes pure state $|\psi\rangle$ in Hilbert space $\mathcal{H}^{(n)}$ (with $\log \dim \mathcal{H}^{(n)} = nR$) in an n-letter codeword which is sent to Bob with n uses of the noisy channel $\mathcal{N}$. Bob applies a decoding map to the n-letter signal he receives obtaining ρ, which has fidelity with the input state $\langle\psi|\rho|\psi\rangle \geq 1-\varepsilon$

We say rate R is achievable if there is a sequence of codes with rate at least R such that $\varepsilon \to 0$ as $n \to \infty$. The quantum capacity $Q(\mathcal{N})$ is the supremum of achievable rates

There is a regularized formula for $Q(\mathcal{N})$.

The channel $\mathcal{N}^{A\to B}$ can be realized by an isometry $\mathcal{N}^{A\to BE}$ where E is an environment.

For any density operator ρ_A on A consider its purification ψ_{RA}, where R is a reference system

This is mapped by the channel to a pure state ϕ_{RBE} of reference system R, environment E, and Bob's system B.

We define the "one-shot" quantum capacity as

$$Q_1(\mathcal{N}) = \max_{\rho_A} \left(-H(R|B) \right)$$

The quantity $-H(R|B) = H(B) - H(RB)$ can also be expressed as $-H(R|B) = H(B) - H(E)$ since the state ϕ_{RBE} is pure. It is important enough to have its own name:

$$I_c(R \rangle B) = -H(R|B) = H(B) - H(E)$$

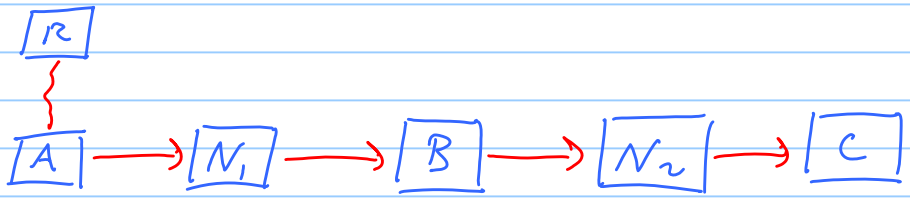
is the "coherent information" from R to B. For a classical channel $H(R|B)$ is always nonnegative and coherent info is never positive. In the quantum setting $I_c(R \rangle B) > 0$ means that the ref. system R has a stronger quantum correlation with B than with the environment E.

The quantum capacity is the regularized quantity

$$Q(N) = \lim_{n \to \infty} \max_{A^n} \frac{1}{n} I_c(R^n \rangle B^n)$$

Here, too, unfortunately, coherent info is superadditive, and we don't know how to express $Q(N)$ as a single-letter formula.

We can understand, though, why coherent info provides an upper bound on quantum capacity



Consider composing two channels N_1 and N_2. Monotonicity of mutual information implies

$$I(R;A) \geq I(R;B) \geq I(R;C)$$

or $H(R) - H(R|A) \geq H(R) - H(R|B) \geq H(R) - H(R|C)$.

since $H(R)$ is unchanged by the channels, this becomes

$$I_c(R\rangle A) \geq I_c(R\rangle B) \geq I_c(R\rangle C)$$

This "quantum data-processing inequality" identifies coherent info as a quantity that cannot be increased by a quantum channel.

But now suppose that the first channel is the noisy channel $\mathcal{N}^{A\to B}$ while the second channel is Bob's decoding map $\mathcal{D}^{B\to C}$. Suppose that ρ_A is maximally mixed on Alice's codespace, so that

$$H(\tilde{A}^n) = H(\tilde{R}^n) = nR \quad \text{where } R \text{ is the code rate}$$

and $H(\tilde{R}^n|\tilde{A}^n) = H(\tilde{A}^n) - H(\tilde{A}^n\tilde{R}^n) = H(\tilde{A}^n) = H(\tilde{R}^n)$

(since the state of $\tilde{A}^n\tilde{R}^n$ is pure). If Bob's recovery is perfect then the state of $\tilde{C}^n\tilde{R}^n$ is also maximally entangled, and

$$I_c(\tilde{R}^n\rangle\tilde{C}^n) = H(\tilde{R}^n) = I_c(\tilde{R}^n\rangle\tilde{A}^n)$$

But therefore the data processing inequality implies

$$I_c(\tilde{R}^n\rangle\tilde{B}^n) = I_c(\tilde{R}^n\rangle\tilde{A}^n) = H(\tilde{R}^n)$$

Since the coherent information is

$$I_c(\tilde{R}^n\rangle\tilde{B}^n) = H(\tilde{B}^n) - H(\tilde{E}^n)$$

and $H(\tilde{B}^n) = H(\tilde{R}^n\tilde{E}^n)$ because the state of $\tilde{R}^n\tilde{B}^n\tilde{E}^n$ is pure, this condition becomes

$$I_c(\tilde{R}^n \rangle \tilde{B}^n) = H(\tilde{R}^n) \Longrightarrow$$
$$H(\tilde{R}^n \tilde{E}^n) = H(\tilde{R}^n) + H(\tilde{E}^n)$$

– correctability means that there is no correlation between R and E (the environment "knows nothing" about which codeword was sent).

Since perfect decoding fidelity means

$$R = \frac{1}{n} I_c(\tilde{R}^n \rangle \tilde{B}^n)$$

we have the bound on achievable rates (and hence capacity)

$$Q(N) \leq \lim_{n\to\infty} \max_{A^n} \frac{1}{n} I_c(R^n \rangle B^n)$$

as we claimed.

Now we would like to argue that $I_c(R\rangle B) = H(B) - H(E)$ is an achievable rate, by using "random quantum codes." We need to explain

1. What is a "random quantum code?
2. That $H(RE) \cong H(R) + H(E)$ is sufficient as well as necessary for (approximate) correctability
3. That $H(RE) \cong H(R) + H(E)$ is true for random codes with rate less than coherent information (except for negligible corrections).

Ph/CS 219
2 March 2011
Jointly typical sequences

Here is a little more detail about the proof of the Noisy Channel Coding Theorem, specifically the proof that the mutual information is an achievable rate. (Based on Chapter 8 of Cover and Thomas.)

A noisy classical channel is characterized by the conditional probability function p(y|x), the probability that the letter y is received when the letter x is sent. We consider using the channel n times to send n letters. We use a code with rate R; that is we send one of 2^{nR} n-letter messages, so that we attempt to convey nR bits of information in n uses of the channel. We say that the rate R is achievable if there is a sequence of codes with rate R such that the probability of of a decoding error approaches zero as n -> infinity. The capacity of the channel is the supremum of achievable rates.

Following Shannon, we consider constructing an n-letter code by generating 2^{nR} codewords, each time sampling from an i.i.d. probability distribution, in which the letter x is selected with probability p(x). For any such code, we consider a codeword selected uniformly at random from among the 2^{nR} possible codewords. We would like to obtain an upper bound on the probability of a decoding error when this codeword is sent through n uses of the channel. To do that we have to choose and analyze a decoding procedure.

Our decoding procedure is "jointly typical decoding". When a correlated probability distribution p(x,y) is sampled n times to generate strings

$$(X_1 X_2 \ldots X_n, Y_1 Y_2 \ldots Y_n) \equiv (\bar{X}, \bar{Y}),$$

by the law of large numbers we know that for each fixed $\delta, \varepsilon > 0$ we can choose n sufficiently large so that, with probability $> 1-\varepsilon$ $(\bar{X}, \bar{Y})$ is "jointly δ-typical", i.e.

$$2^{-n[H(X)+\delta]} \le p(\bar{x}) \le 2^{-n[H(X)-\delta]},$$
$$2^{-n[H(Y)+\delta]} \le p(\bar{y}) \le 2^{-n[H(Y)-\delta]},$$
$$2^{-n[H(XY)+\delta]} \le p(\bar{x},\bar{y}) \le 2^{-n[H(XY)-\delta]}.$$

The number N_{typ} of jointly typical sequences satisfies

$$1 \ge \sum_{typ} p(\bar{x},\bar{y}) \ge N_{typ}\, 2^{-n[H(XY)+\delta]}$$
$$\Rightarrow \quad N_{typ} \le 2^{n[H(XY)+\delta]}$$

When Bob receives $\bar{y}$, then if there is unique codeword $\bar{x}$ jointly typical with $\bar{y}$ he decodes $\bar{y}$ as $\bar{x}$. Otherwise he decodes in an arbitrary way.

A decoding error occurs if either of the following happen:

① The sent and received messages are not jointly typical. This occurs with probability $\leq \varepsilon$.

② There is a codeword $\bar{x}$ other than the one sent which is jointly typical with the received message $\bar{y}$.

Suppose that $\bar{x}_1$ was actually sent and $\bar{y}_1$ received, and let $\bar{x}_2$ be another codeword different than $\bar{x}_1$. What is the probability that $\bar{x}_2$ and $\bar{y}_1$ are jointly typical?

Because $\bar{x}_2$ and $\bar{y}_1$ were determined by sampling independently, they are uncorrelated: the probability that $\bar{x}_2$ and $\bar{y}_1$ were generated factorizes into $p(\bar{x}_2)\, p(\bar{y}_1)$ (the product of the marginal distributions for $\bar{x}$ and $\bar{y}$). The probability of joint typicality is

$$\sum_{(\bar{x}_2, \bar{y}_1)\ \text{typ}} p(\bar{x}_2)\, p(\bar{y}_1) \leq 2^{n H[(X,Y)+\delta]}\, 2^{-n[H(X)-\delta]}\, 2^{-n[H(Y)-\delta]}$$

upper bound on N_{typ} ↗ upper bound on prob. of typical $\bar{x}_2$ ↗ upper bound on prob. of typical $\bar{y}_1$ ↗

$$\leq 2^{-n[I(X;Y)-3\delta]}$$

There are $(2^{nR}-1)$ codewords other than $\bar{x}_1$ that might have been sent. So, averaged over codes and codewords,

$$\text{Prob. of decoding error} \leq \varepsilon + (2^{nR}-1)\, 2^{-n[I(X;Y)-3\delta]}$$

$$\leq \varepsilon + 2^{-n[I-R-3\delta]} \longrightarrow 0 \quad \text{as } n \to \infty$$

for any $R < I$

Slepian-Wolf coding

In Sec. 5.1.2 of the lecture notes, it is claimed that if a joint distribution p(x,y) is sampled n times, where Alice receives n-letter message x and Bob receives n-letter message y, then Alice can send nH(X|Y) bits to Bob, enabling Bob to determine x with high asymptotic success probability. Here we explain in more detail the coding scheme that Alice and Bob use to achieve this. It is a special case of "Slepian-Wolf coding" (Cover and Thomas Sec. 14.4).

Alice sorts all possible n-letter messages into 2^{nR} bins which are selected uniformly at random. The choice of bins is known to both Alice and Bob. Alice sends to Bob the nR bits that identify the bin that contains her message x. Thus Bob knows both y and the bin; he decodes y as x if x is the unique message in this bin that is jointly typical with y. Otherwise he choses an arbitrary decoding.

A decoding error occurs if
(1) The Alice's message x and Bob's message y are not jointly typical. This occurs with probability no larger than epsilon.

If $(\bar{x},\bar{y})$ are jointly typical, then

$$p(\bar{x}|\bar{y}) = \frac{p(\bar{x},\bar{y})}{p(\bar{y})} \geq \frac{2^{-n[H(XY)+\delta]}}{2^{-n[H(Y)-\delta]}} = 2^{-n[H(X|Y)+2\delta]}$$

If y is typical, let $S(X|\bar{y})$ denote the set of $\bar{x}$ that are jointly typical with $\bar{y}$. Then

$$1 \geq \sum_{\bar{x}\in S(X|y)} p(\bar{x}|\bar{y}) \geq |S(X|y)|\, 2^{-n[H(X|Y)+2\delta]}$$

The no. of elements of $S(X|y)$

$$\Longrightarrow |S(X|y)| \leq 2^{n[H(X|Y)+2\delta]}$$

Because the bins are chosen uniformly at random, each $\bar{x}$ is contained in a particular specified bin with probability 2^{-nR}. The probability that $\bar{x}$ is in the bin containing Alice's message by accident is

$$\leq 2^{-nR}|S(X|y)|$$

$$\leq 2^{-n[R-H(X|Y)-2\delta]} \longrightarrow 0 \text{ as } n\to\infty$$

for $R > H(X|Y)$.

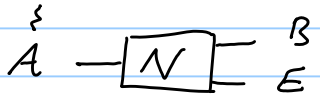
Coding in which Alice sends $H(X|Y)$ bits per letter is achievable. (If ave. over codes has decoding error probability $< \varepsilon$, then there is a particular code with error probability $< \varepsilon$.)

Ph/CS 219
Quantum Channel Capacities
7 March 2011

Note Title 3/5/2011

Last time we introduced the concept of "coherent information" and noted its relevance to sending quantum information through a noisy quantum channel. A channel $\mathcal{N}^{A\to B}$ has a dilation, the isometry $\mathcal{N}^{A\to BE}$

Suppose that input density operator ρ_A is purified by reference system R. Sending A through the channel prepares the tripartite pure state ϕ^{RBE}



The coherent information from R to B for channel $\mathcal{N}$ and input ρ_A is

$$I_c(R\rangle B) = -H(R|B) = H(B) - H(E),$$

it does not depend on the choice of purification, since a unitary on R does not change $H(B)$ or $H(E)$. It can also be expressed as

$$I_c(R\rangle B) = \tfrac{1}{2}\left[I(R;B) - I(R;E)\right],$$

since
$$I(R;B) = H(R) + H(B) - H(RB) = H(R) + H(B) - H(E)$$
$$I(R;E) = H(R) + H(E) - H(RE) = H(R) + H(E) - H(B),$$

and hence it quantifies how much stronger the correlation of R is with B than with E.

If the signal transmitted through the channel can be perfectly corrected, then Bob can apply a decoding map with dilation $\mathcal{D}^{B\to \hat{B}B'}$ such that

$$\phi^{RA} \xrightarrow{\mathcal{N}} \phi^{RBE} \xrightarrow{\mathcal{D}} \phi^{R\hat{B}}\otimes\psi^{B'E}$$

We argued that Bob can decode perfectly only if

$$H(R|B) = I_c(R\rangle B) \quad \text{or} \quad H(RE) = H(R) + H(E)$$

That is, for perfect correctability we

require that the state of RE is a product state — R and E are uncorrelated, or "decoupled."

By considering n uses of the channel, and choosing R to purify the maximally mixed state on the code space, we concluded that the regularized coherent information is an upper bound on the achievable rate for high-fidelity quantum communication:

$$Q(N) \leq \lim_{n\to\infty} \max_{A^n} \frac{1}{n} I_c(R^n \rangle B^n).$$

Conversely, if R is maximally entangled with the code space, decoupling of RE suffices to ensure that any state in the code space can be perfectly decoded. If ϕ^{RBE} is the purification of RE density operator $\sigma^{RE} = \sigma^R \otimes \sigma^E$, then we can split B into two subsystems $B = \hat{B}B'$ such that $\hat{B}$ purifies σ^R and B' purifies σ^E; i.e.

$$\phi^{RBE} = \Phi^{R\hat{B}} \otimes \psi^{B'E}$$

where $\Phi^{R\hat{B}}$ is max. entangled,

and therefore Bob can construct a decoding map $\mathcal{D}^{B\to\hat{B}E'}$ that extracts Alice's logical state in the subsystem $\hat{B}$.

Furthermore, approximate decoupling of RE suffices for approximate correctability.

Recall that the fidelity of density operators is defined as

$$F(\rho,\sigma) = \left(\mathrm{tr}\sqrt{\rho^{1/2}\sigma\rho^{1/2}}\right)^2 = \|\sqrt{\rho}\sqrt{\sigma}\|_1^2$$

and is related to L^1 distance between ρ and σ by

$$F(\rho,\sigma) \geq 1 - \|\rho-\sigma\|_1 \quad \text{(see Appendix B)}$$

Also, if $|\psi_\rho\rangle$ is a purification of ρ, then

$$F(\rho,\sigma) = \max |\langle \psi_\sigma | \psi_\rho \rangle|^2 \qquad (\text{"Uhlmann's Theorem"})$$

(where the max is over all possible purifications of σ). So, suppose ρ^{RE} is close to a product state:

$$\| \sigma^{RE} - \sigma^R_{max} \otimes \sigma^E \|_1 \leq \varepsilon$$

(where σ^R_{max} is the maximally mixed state on R). Then σ^{RE} has a purification that has large overlap with the purification of $\sigma^R_{max} \otimes \sigma^E$:

$$|\langle \tilde{\phi}^{RBE} | \phi^{RBE} \rangle|^2 \geq 1 - \varepsilon$$

where $|\phi^{RBE}\rangle$ is the purification of σ^{RE} and $|\tilde{\phi}\rangle^{RBE} = |\Phi\rangle^{R\hat{B}} \otimes |\psi\rangle^{B'E}$ is the purification of $\sigma^R_{max} \otimes \sigma^E$

When we trace out a subsystem, fidelity is monotonic (states cannot become easier to distinguish), so applying the decoding map $\mathcal{D}^{B \to \hat{B}}$ to $|\phi^{RBE}\rangle$ yields $F(\Phi^{R\hat{B}}, \mathcal{D}^{B\to\hat{B}}(\sigma^{RB})) \geq 1 - \varepsilon$

In other words, after decoding the density operator of R and Bob's decoded subsystem $\hat{B}$ is $\sigma^{R\hat{B}}$ where

$$\langle \Phi^{R\hat{B}} | \sigma^{R\hat{B}} | \Phi^{R\hat{B}} \rangle \geq 1 - \| \sigma^{RE} - \sigma^R_{max} \otimes \sigma^E \|_1 .$$

We conclude: approx. decoupling implies approx correctability.

Aside: Proof of Uhlmann's Thm

Purification of ρ can be expressed as

$$\sum_a \sqrt{\lambda_a}\, |e_a\rangle \otimes |f_a\rangle = (\rho^{\frac{1}{2}} \otimes I)\, |\tilde{\Phi}\rangle \text{ where } |\tilde{\Phi}\rangle = \sum_a |e_a\rangle \otimes |f_a\rangle$$

and $\rho = \sum_a \lambda_a |e_a\rangle\langle e_a|$, and an arbitrary purification of σ is

$$|\psi_\sigma\rangle = \sum_i \sqrt{\eta_i}\, |g_i\rangle \otimes |h_i\rangle = (\sigma^{\frac{1}{2}} \otimes I)\, |\tilde{\Psi}\rangle \quad \left(\text{where } |\tilde{\Psi}\rangle = \sum_i |g_i\rangle \otimes |h_i\rangle\right)$$

$$= (\sigma^{\frac{1}{2}} \otimes I)(V \otimes W^T)\, |\tilde{\Phi}\rangle = \sigma^{\frac{1}{2}} V W \otimes I)\, |\tilde{\Phi}\rangle$$

where $\sigma = \sum_i \eta_i |g_i\rangle\langle g_i|$ and V, W are unitary.

Thus $\langle \psi_\sigma | \psi_\rho \rangle = \langle \tilde{\Phi} | (U^\dagger \otimes I)\, \sigma^{\frac{1}{2}} \rho^{\frac{1}{2}} | \tilde{\Phi} \rangle$

(where $U = VW$) $= \mathrm{tr}\left(U^\dagger \sigma^{\frac{1}{2}} \rho^{\frac{1}{2}}\right)$

Using the polar decomp $A = U' \sqrt{A^\dagger A}$ applied to $A = \sigma^{\frac{1}{2}} \rho^{\frac{1}{2}}$

this is $\langle \psi_\sigma | \psi_\rho \rangle = \mathrm{tr}\left(U^\dagger U' \sqrt{\rho^{\frac{1}{2}} \sigma \rho^{\frac{1}{2}}}\right)$

whose modulus is maximized by choosing $U = U'$ so that

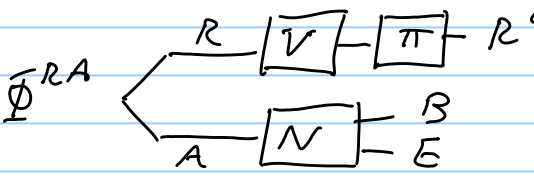
$$\max |\langle \psi_\sigma | \psi_\rho \rangle| = \left(\mathrm{tr} \sqrt{\rho^{\frac{1}{2}} \sigma \rho^{\frac{1}{2}}}\right)^2 \quad \text{as claimed.}$$

Monotonicity is a corollary: $F(\rho_{AB}, \sigma_{AB}) \leq F(\rho_A, \sigma_A)$, because any purifications of ρ_{AB} and σ_{AB} are also purifications of ρ_A and σ_A.

Achievability of Coherent Info

To show that coherent info is an achievable rate, we use a random quantum code. When using the channel n times, chose a random subspace of A^n as input to $(\mathcal{N}^{A \to B})^{\otimes n}$.

That is, consider



π projects R to a fixed subspace R' and V is a unitary on R, so V determines what subspace is projected.

Φ^{RA} is a maximally entangled state of RA, so R' purifies the maximally mixed state on a code space determined by V.

Now we can average over V. One can show that for any state σ^{RE} on RE, if R' is random

subspace of R determined by V, then

$$\left(\int dV \; \| \sigma^{R'E}(V) - \sigma^{R'}_{max} \otimes \sigma^{E} \|_1 \right)^2 \leq |R'E| \, tr\left(\sigma^{RE}\right)^2$$

Here $\bar{V}$ is the normalized unitarily-invariant (Haar) measure on the unitary group acting on R.

In the case where we used the channel n times, the state on B^n is nearly maximally mixed on a typical subspace of dimension $|B^n| \cong 2^{nH(B)}$, the state on E^n is nearly maximally mixed on a typical subspace of dimension $|E^n| \cong 2^{nH(E)}$, and the state on R^n is nearly maximally mixed on a typical subspace of dimension $|R^n| \cong 2^{nH(R)}$. We apply the encoding unitary to this typical subspace R^n before projecting onto R' with $|R'| = 2^{n(Rate)}$,

where "Rate" is the rate of the code in qubits per use of the channel.

Suppressing the small δ in the estimate of the dimension, we estimate $tr\left(\rho^{RE}\right)^2 = tr\left(\rho^{B}\right)^2 \cong \frac{1}{|B^n|}$

and we conclude that, when we average over codes, the deviation of $R'E$ from a product state is suppressed for

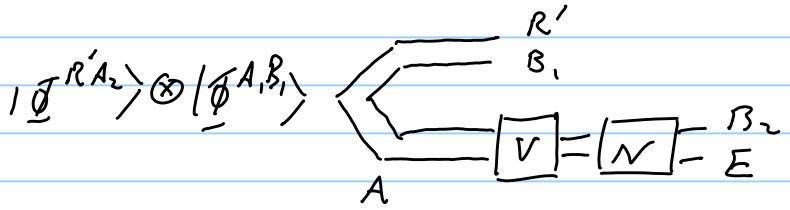
$$\frac{|R'|\,|E^n|}{|B^n|} \approx \frac{2^{n(Rate)}\, 2^{nH(E)}}{2^{nH(B)}} \ll 1$$

or $\quad Rate < H(B) - H(E) = I_c(R \rangle B).$

Since decoupling is well satisfied when we average over the choice of the encoding unitary V, then RE decouples well for some particular V (and in fact for a typical V).

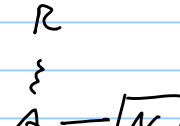
Father Protocol: Entanglement assisted quantum communication

It is also instructive to estimate the rate for entanglement-assisted quantum communication. Now the sender A and receiver B share a supply of entangled qubits that are consumed during the protocol.



In the i.i.d version of the protocol (many uses of the channel) Alice and Bob share a maximally entangled state $\phi^{A_1B_1}$ and Alice's input qubits A_2 are maximally entangled with reference system R' (the state ϕ^{RA_2}). To encode Alice applies a typical unitary V that acts collectively on the input system A_2 and her half of the entangled qubits. Bob's decoding map can act collectively on his half of the shared entanglement and the output he receives through the (uses of) the noisy channel. For Bob to be able to decode successfully, it suffices that $R'E$ decouple.

This protocol for entanglement-assisted quantum communication is called the "Father protocol" because it has a variety of interesting "children" that can be derived as consequences.



Recall again that for any input density operator ρ_A, we may consider its purification ϕ^{RA} and the pure state ϕ^{RBE} resulting from sending A through the channel with dilation $N^{A\to BE}$. The Father resource inequality expresses an achievable rate for the quantum communication in the Father protocol, and also the cost in Bell pairs for achieving that rate, in terms of properties of ϕ^{RBE}. Namely

$$\langle \mathcal{N}^{A\to B} : \rho_A \rangle + \tfrac{1}{2} I(R;E)[qq] \geq \tfrac{1}{2} I(R;B)[q\to q]$$

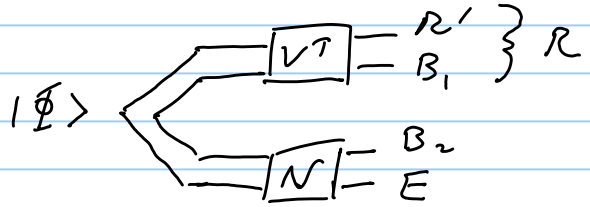
This means that, asymptotically, by using the noisy channel n times, $\frac{n}{2} I(R;B) - o(n)$ qubits can be sent from A to B with high fidelity while consuming $\frac{n}{2} I(R;E) + o(n)$ ebits of entanglement. (Here $o(n)$ means a quantity increasing more slowly than linearly in n.) The entropic quantities depend on the density operator ρ_A, and the resource inequality expresses a task that can be achieved for any ρ_A, so we are free to choose ρ_A that optimizes the rate.

To help you remember the Father inequality, note that $I(R,E)$ quantifies something bad — the noise. The higher $I(R;E)$ is, the more entanglement we need to do something useful. On the other hand, $I(R;B)$ quantifies something good — the correlation that survives transmission through the noisy channel. The higher $I(R;B)$ is, the higher the rate of quantum communication. (But the factor $1/2$ you will just need to remember.)

We can relate the Father protocol to an even more primitive task called the "mother protocol" or "quantum state transfer" protocol. Recalling that

$$(I\otimes V)|\Phi\rangle = (V^T\otimes I)|\Phi\rangle$$

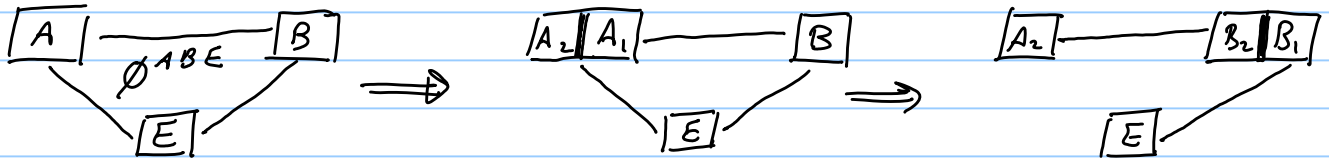
when $|\Phi\rangle$ is maximally entangled, the father transforms into



Now there is a tripartite state ϕ^{RB_2E} where Roy holds R

Roy divides R into subsystems $R = R'B_1$ (where the decomposition depends on $\bar{V}$); he keeps R' and passes B_1 to Bob. If $|B_1|$ is large enough, then $R'E$ decouples, which means that the system maximally entangled with R' can be recovered by Bob after decoding. This means that the corresponding father protocol conveys $\log|R'|$ qubits from Alice to Bob while consuming $\log|B_1|$ ebits of entanglement

Changing Roy's name to Alice, and relabeling the subsystems, the mother protocol can be described this way. Alice, Bob, and Eve share the tripartite pure state ϕ^{ABE}. Alice divides her system into subsystems, $A = A_1A_2$; she keeps A_2 and sends A_1 to Bob. Her goal is to send enough qubits to Bob so that what she holds is no longer correlated with Eve. At that point, the purification of Eve's state is entirely in Bob's hands and Bob also holds the purification of A_2; i.e. Bob's system at the end of the protocol has decomposition $A_1B = B_1B_2$ where B_1 purifies E and B_2 purifies A_2.



In the i.i.d. version of the mother A, B, E share many identical copies $(\phi^{ABE})^{\otimes n}$. Alice Schumacher compresses to a typical subspace of dimension $n(H(A) + o(n))$ and then sends a random subsystem A_1 to Bob. Bob decodes by dividing his system into B_1B_2.

The mother resource inequality expresses how many qubits of quantum communication from A to B suffice to decouple A_2 and E, and how many ebits of entanglement reside in A_2B_2 when the protocol ends:

$$\langle \phi^{ABE} \rangle + \frac{1}{2} I(A;E)[q \to q] \geq \frac{1}{2} I(A;B)[qq] + \langle \phi'^{B,E} \rangle$$

That is, $\frac{n}{2}[I(A;E) + o(1)]$ qubits of communication decouple A and E (each qubit sent reduces the mutual information by two bits). Meanwhile A and B harvest $\frac{n}{2}[I(A;B) - o(1)]$ ebits of entanglement. This mother protocol is "dual" to the father protocol - now quantum communicatim is consumed and quantum entanglement is achieved, rather than the other way around. $I(A;E)$ quantifies the noise in the entanglement that $A+B$ share at the beginning of the protocol, and $I(A;B)$ quantifies the correlation between $A+B$ at the beginning.

The mother can be viewed as a generalization of the entanglement concentration protocol discussed earlier, extended in 3 ways:

1. The initial state shared by $A+B$ can be mixed rather than pure.
2. The communicatim from A to B is quantum rather than classical.
3. We quantify the amount of communicatim required.

Note also that if the state of AE is pure, the mother becomes Schumacher compression: Alice sends $\frac{n}{2} I(A;E) = nH(A)$ qubits to Bob and $I(A;B) = 0$. Eve can measure E to realize ρ_A as an ensemble of pure states.

In additim, as we have seen, the mother resource inequality implies the father, if we think of the communication from Alice to Bob in the mother as the offloading of part of R from Roy to Bob in the father, so that the amount of quantum communication in the mother is the quantum entanglement consumed by the father. Noting that

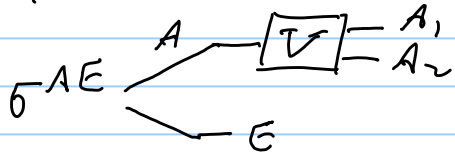
$$H(R) = \frac{1}{2} I(R;B) + \frac{1}{2} I(R;E),$$

we see that if Roy sends $\frac{n}{2} I(R;E) + o(n)$ qubits to Bob, he retains a reference system R'

with $\frac{n}{2} I(R;B) - o(n)$ qubits, which becomes the number of qubits in the code used in the father protocol, while the $\frac{n}{2} I(R;E) + o(n)$ qubits sent by Roy in the mother becomes the number of ebits consumed in the father.

Achievable rate in mother protocol

Consider an arbitrary mixed state σ^{AE} of AE. Consider a fixed decomposition into subsystems $A = A_1 A_2$. Apply a unitary V to A before discarding A_1 to obtain marginal state $\sigma^{A_2 E}(V)$.



The "decoupling inequality" expresses how close $A_2 E$ is to a product state when we average V over unitaries acting on A with respect to Haar measure:

$$\left(\int dV \, \| \sigma^{A_2 E}(V) - \sigma_{\max}^{A_2} \otimes \sigma^{E} \|_1 \right)^2 \leq \frac{|A| \cdot |E|}{|A_1|^2} \operatorname{tr}(\sigma^{AE})^2$$

(where $\sigma_{\max}^{A_2}$ is the maximally entangled state on A_2).

This generalizes the result found in a homework exercise, which concerned the case where E is trivial and σ^A is pure; there you derived:

$$\left(\int dV \, \| \sigma^{A_2}(V) - \sigma_{\max}^{A_2} \|_1 \right)^2 \leq \frac{|A_2|}{|A_1|} = \frac{|A|}{|A_1|^2}.$$

($\sigma^{A_2}(V)$ nearly maximally mixed for $|A_2| \ll |A_1|$.)

In the i.i.d. version of the mother A becomes (after Alice performs Schumacher compression) the typical subspace of A^n, E the typical subspace of E^n, AE the typical subspace of $(AE)^n$. Since AE is

nearly maximally mixed on space of dim $\simeq 2^{nH(AE)}$ we have $\mathrm{tr}(\rho^{AE})^2 \simeq 2^{-nH(AE)}$. Therefore, when we average over V, the state on A_2E is nearly a product state provided

$$\frac{1}{|A_1|^2} 2^{nH(A)} 2^{nH(E)} 2^{-nH(AE)} << 1$$

or $$|A_1|^2 >> 2^{nI(A;E)}.$$

It suffices then for Alice to send

$$\log |A_1| = \frac{n}{2} I(A;E) + o(n)$$

qubits to Bob. And since

$$H(A) = \frac{1}{2}[I(A;E) + I(A;B)] \quad \text{(because } \phi^{ABE} \text{ is pure)}$$

Alice retains

$$\log |A_2| = \frac{n}{2} I(A;B) - o(n)$$

qubits. Since these are nearly maximally mixed and uncorrelated with E, Alice's retained qubits are nearly maximally entangled with a subsystem of Bob's qubits; Alice and Bob share $\frac{n}{2} I(A;B) - o(n)$ ebits. This proves the mother resource inequality. (It works when we average over V, and therefore for some particular V — in fact for typical V.)

There is also an $o(1)$ contribution to the RHS of the decoupling inequality from portion of $(\phi^{ABE})^{\otimes n}$ that lies outside typical subspace, which vanishes in limit $n \to \infty$.

The proof of the decoupling inequality is in Appendix A. Note that a simple heuristic dimension counting argument shows that it is plausible, at least in the i.i.d. case that is relevant for the asymptotic achievability result. Suppose that the state on AE is maximally mixed on a subspace of dim $|B|$, i.e., a uniform mixture of $|B|$ mutually orthogonal pure states. Then we trace out A_1. But for $|A_1| << |A_2E|$, we expect that each of the $|B|$ states in the ensemble realizing ρ^{AE} is likely to be nearly maximally mixed

on A_1; thus for each of these $|B|$ states, tracing out A_1 generates a density operator on A_2E which is a nearly uniform mixture of $|A_1|$ mutually orthogonal states. Furthermore, as long as $|A_1B| \ll |A_2E|$, all of the $|A_1B|$ states are likely to be nearly mutually orthogonal — tracing out A_1 produces a nearly uniform density operator with rank $\approx |A_1B|$. Once $|A_1|$ is large enough though, the rank $|A_1B|$ matches the dimension of A_2E, so that the state on A_2E is maximally mixed and in particular is a product state. This occurs for

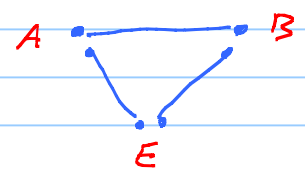
$$|A_1| \cdot |B| \approx |A_2| \cdot |E| = \frac{|A| \cdot |E|}{|A_1|}$$

or $$|A_1|^2 \approx \frac{|A| \cdot |E|}{|B|}$$

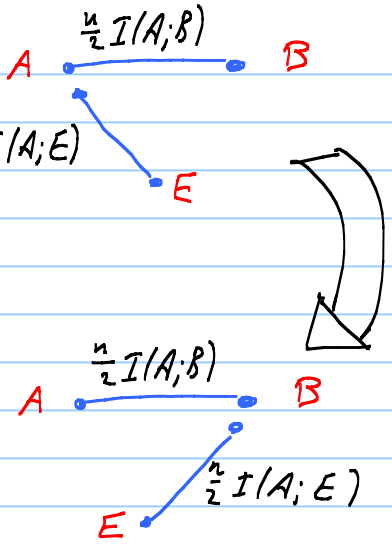
$$\approx \frac{2^{nH(A)} 2^{nH(E)}}{2^{nH(AE)}} = 2^{nI(A;E)},$$

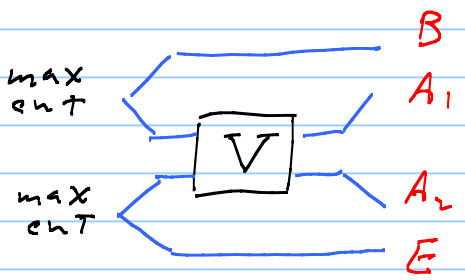
reproducing the conclusion we inferred from the decoupling inequality.

We may think about the state transfer ("mother") protocol this way: In the iid case, where Alice, Bob, and Eve share n copies of ϕ^{ABE}, the typical subspaces of A^n, B^n, E^n are all nearly maximally mixed. Therefore A is maximally entangled with a subsystem of BE, B with a subsystem of AE, and E with a subsystem of AB. Consider the "trivial" case, in which A divides into two subsystems, one entangled with B and the other with E. Similarly, B divides into the subsystem entangled with A and a complementary subsystem entangled with E, and E divides into two subsystems, one entangled with A, the other with B.



For the state transfer task, the entanglement shared by BE is not relevant, so ignore it. Alice shares $\frac{n}{2}I(A;B)$ ebits with Bob and $\frac{n}{2}I(A;E)$ ebits with Eve. If Alice knows which of here $nH(A)$ qubits are entangled with E, she can send those $\frac{n}{2}I(A;E)$ qubits to Bob and so "decouple" from Eve.





But what is remarkable is that Alice does not need to know! If K of her n qubits are maximally entangled with E and n-K are max. entangled with B, she can send K+ constant randomly chosen qubits to Bob (the subsystem A_1); then A_2E will be almost a product state, and a subsystem of A_1B will be almost maximally entangled with E. And A_2, since it is max mixed and decoupled from E, is also max. entangled with A, B. The protocol is obviously optimal asymptotically, since K ebits of entanglement between E and A, B cannot be established by sending fewer than K qubits.

If, for example, E, B are $\frac{n}{2}$-qubit systems, each maximally entangled with a subsystem of A's n qubits, Alice can select $\frac{n}{2}$ + constant random qubits and send them to either E or B. Either way, what Alice retains decouples from the other system!

As usual, this random coding argument establishes the existence of a protocol that achieves the mother resource inequality, without exhibiting any particular protocol. Furthermore, most unitary transformations on n qubits can be approximated accurately only by a quantum circuit of size exponential in n, so the argument based on averaging over unitaries does not ensure there exists a protocol that can be executed efficiently.

It turns out, though, that the decoupling argument works if we average over the n-qubit Clifford group, and these transformations can be realized by poly-sized circuits. So selecting which subsystem to send to Bob can be done efficiently — there is a stabilizer code that does the job! Also, there are poly-sized circuits that achieve Schumacher compression. So Alice's encoding can be done efficiently in the mother protocol. (Bob's decoding is easy too, since a stabilizer code, encoded by the Clifford circuit, is used to correct erasure.)

Children of the Father

We can derive a further consequence by combining the father resource inequality

$$\text{Father: } \langle \mathcal{N}^{A\to B}, \rho_A \rangle + \tfrac{1}{2} I(R;E)[q\to q] \geq \tfrac{1}{2} I(R;B)[q\to q]$$

with the superdense coding inequality

$$\text{SD: } \quad [q\to q] + [q\to q] \geq 2[c\to c]$$

(we use one qubit of quantum comm. and one ebit to achieve 2 bits of classical comm.)

Suppose we use the $\frac{1}{2} I(R;B)$ qubits of $[q\to q]$ and an additional $\frac{1}{2} I(R;B)$ ebits to achieve $I(R;B)$ bits of $[c\to c]$. Because

$$\frac{1}{2} I(R;E) + \frac{1}{2} I(R;B) = H(R),$$

we conclude

$$\langle N^{A\to B} : \rho_A \rangle + H(R)[qq] \geq I(R;B)[c\to c],$$

which establishes an achievable rate for entanglement-assisted classical communication.

We may define $C_E(N)$ as the supremum of achievable rates per use of the channel for sending classical info reliably over the noisy quantum channel, if entanglement can be consumed at zero cost. This = "entanglement-assisted classical capacity" of the quantum channel thus satisfies

$$C_E(N) \geq \max_{\rho_A} I(R;B)$$

In this case, there is a matching upper bound, and thus the inequality is actually an equality. In this case, therefore, we have a single-letter formula and the cost of the task is fully understood. Furthermore, the resource inequality tells us how much entanglement consumption suffices to attain the capacity.

We can derive another consequence of the father by using some of the quantum communication generated by the father to repay the entanglement that was borrowed to activate (i.e. catalyze) the father protocol.

$$[q\to q] \geq [qq] \implies \frac{1}{2} I(R;E)[q\to q] \geq \frac{1}{2} I(R;E)[qq].$$

After replacing the entanglement consumed, the net amount of quantum communication achieved per use of the channel is

$$\frac{1}{2} I(R;B) - \frac{1}{2} I(R;E) = H(B) - H(E) = I_C(R\rangle B).$$

we have derived the achievability result

$$\langle \mathcal{N}^{A\to B} : \rho_A \rangle \geq I_c(R\rangle B)[q\to q],$$

at least in this catalyzed setting, and the same rate can also be achieved without any initial supply of entanglement. Together with the upper bound (derived in the homework?) we obtain a regularized formula for quantum capacity:

$$Q(\mathcal{N}) = \lim_{n\to\infty} \max_{\rho_A^{in}} \frac{1}{n} I_c(R^n\rangle B^n)$$

unfortunately, though, since the coherent information can be superadditive, we don't know how to reduce this expression to a single-letter formula for the quantum capacity.

Children of the mother

We obtain a useful consequence of the mother resource inequality

$$\text{Mother: } \langle \phi^{ABE}\rangle + \tfrac{1}{2} I(A;E)[q\to q] \geq \tfrac{1}{2} I(A;B)[qq] + \langle \phi'^{B,E}\rangle$$

by combining with the teleportation resource inequality

$$\text{TP: } [qq] + 2[c\to c] \geq [q\to q]$$

(one qubit can be transmitted by consuming one ebit and sending two bits.)

We can replace the quantum communication in the mother by classical communication if we use $\frac{1}{2} I(A;E)$ ebits generated by the mother, together with $I(A;E)$ bits of classical communication to replace the quantum communication consumed by the mother. Then the net amount of entanglement

generated is $\frac{1}{2} I(A;B) - \frac{1}{2} I(A;E) = I_c[A \rangle B]$,

and we obtain the resource inequality

$$\langle \phi^{ABE} \rangle + I(A;E)[c \to c] \geq I_c(A \rangle B)[qq] + \langle \phi'^{B,E} \rangle,$$

which is called the "Hashing inequality." It quantifies an achievable rate for distilling maximal entanglement from a state shared by A and B using one-way classical communication from A to B. Furthermore, the Hashing inequality tells us how much classical communication suffices.

In the case where the state on AB is pure, $I_c(A \rangle B) = H(A) - H(AB) = H(A)$, and we recover our earlier conclusion concerning entanglement concentration for pure states:

$$\langle \phi^{AB} \rangle \geq H(A)[qq],$$

$H(A)$ ebits can be extracted asymptotically from n copies of ϕ^{AB}. In this case the resource inequality says that the sufficient amount of $[c \to c]$ is $I(A;E) = 0$ — the classical communication required is negligible

State Merging

The state-merging resource inequality answers the question: how much quantum communication is needed from A to B to transfer the purification of E's state shared by AB to a state held solely by B, assuming classical communication from A to B has zero cost. To derive state merging from the mother, we use all

of the entanglement generated by the mother to teleport additional qubits from A to B. Adding

$$TP: \quad \tfrac{1}{2} I(A;B)[q\to q] + I(A;B)[c\to c] \geq \tfrac{1}{2} I(A;B)[q\to q]$$

to the mother inequality, and noting that the net amount of quantum communication consumed is

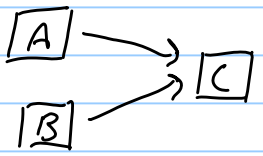
$$\tfrac{1}{2} I(A;E) - \tfrac{1}{2} I(A;B) = H(E) - H(B) = H(AB) - H(B) = H(A|B),$$

we obtain

State Merging: $$\langle \phi^{ABE} \rangle + H(A|B)[q\to q] + I(A;B)[c\to c] \geq \langle \phi'^{B,E} \rangle.$$

state-merging is achieved with an amount of quantum communication given by the conditional entropy $H(A|B)$.

What is the classical version of state merging? If Alice and Bob have correlated classical bits, how many bits does Alice need to send to Bob so that Bob knows what Alice had? The answer is the conditional entropy $H(X|Y)$, which is achieved by what information theorists call "Slepian-Wolf coding". Alice sorts her messages into $2^{n(H(X|Y)+\delta)}$ bins and sends only the label of the bin. With high probability, Bob finds that only one message in that bin is jointly typical with his information.



Similarly, if A and B both send to C: Bob compresses the info from his source to $nH(Y) + o(n)$ letters. Then Alice need send only $nH(X|Y) + o(n)$ letters to C. Together, AB compress their shared information source to $nH(XY)$ letters, the same compression they would have been able to achieve if they were sending from the same location instead of two different locations. Therefore

Slepian-Wolf coding gives a precise operational interpretation to the informal statement that $H(X|Y)$ quantifies Bob's remaining ignorance about XY when he already knows Y.

In the same sense, state merging gives such an operational meaning to conditional entropy in the quantum setting: $H(A|B)$ is the number of qubits Bob needs to receive from Alice in order to possess the purification of system E (if classical communication is for free). The conditional entropy quantifies Bob's "ignorance" about this jointly held purification.

Classically, $H(X|Y)$ is nonnegative, and it is zero if Bob is already certain about XY. But quantumly, $H(A|B)$ can be negative. How can Bob have "negative uncertainty" about AB? If $H(A|B) < 0$ (equivalently $I(A;B) > I(A;E)$), then the mother produces more entanglement than the amount of quantum communication it consumes. In that case, the state merging inequality becomes the hashing inequality

Hashing: $\langle \phi^{ABE} \rangle + I(A;E)[c \to c] \geq -H(A|B)[qq] + \langle \phi'^{BE} \rangle$

Now the state merging has no quantum cost, and AB hold $-H(A|B)$ ebits at the end of the protocol. This shared $[qq]$ they have deposited in the bank can be used for teleportation in future rounds of state merging, reducing the quantum communication cost. The "negative uncertainty" Bob has today can reduce his uncertainty in the quantum communication tasks he will need to perform tomorrow.

Operational meaning of strong subadditivity

The observation that $H(A|B)$ is the quantum communication cost of state merging provides a simple "operational proof" of the strong subadditivity of quantum mutual information. SSA says

$$I(A;BC) = H(A) - H(A|BC) \geq I(A;B) = H(A) - H(A|B)$$

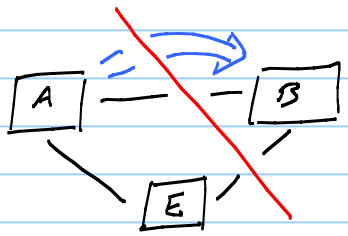
or equivalently:

$$H(A|BC) \leq H(A|B)$$

If $H(A|B)$ is positive, this is the obvious statement that it is no harder to merge A with Bob's system if Bob holds C as well as B.

If $H(A|B)$ is negative, this is the obvious statement that Alice and Bob can distill no less entanglement with one-way classical communication if Bob holds C as well as B.

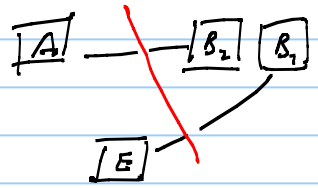
To complete the argument, we need to know that $H(A|B)$ is not only achievable, but also the optimal cost of state merging / hashing. This follows because for a bipartite pure state n qubits of quantum communication from A to B cannot increase the entanglement shared by A and B by more than n ebits.



Consider the entanglement across the cut between B and AE. In the Hashing protocol, the entanglement at the beginning of the protocol is $nH(B)$. At the end E is decoupled from A and has entanglement $nH(E)$ with B. The total entanglement is

$$nH(E) + K \leq nH(B) \quad \text{if } K \text{ ebits are distilled}$$

$$\Rightarrow \frac{K}{n} \leq H(B) - H(E) = -H(A|B)$$



Thus $-H(A|B)$ is the maximal # of ebits that can be distilled per copy.

For state merging, the initial entanglement between B and AE is $nH(B)$. At the end of the protocol, B is entangled with E, so the entanglement across the cut is at least $nH(E)$. To achieve this increase in entanglement the number of qubits sent from A to B must be at least

$$K \geq nH(E) - nH(B) \Rightarrow \frac{K}{n} \geq H(E) - H(B) = H(A|B).$$

At least $H(A|B)$ qubits must be sent per copy.

This argument is a proof of strong subadditivity, since we proved the achievability of state merging and hashing using the decoupling inequality and the theory of typical subspaces, both of which were proved without using strong subadditivity.

Appendix A: The decoupling inequality

We want to show

$$\left(\int dU \, \| \sigma^{A_2 E}(U) - \sigma^{A_2}_{max} \otimes \sigma^{E} \|_1 \right)^2 \leq \frac{|AE|}{|A_1|^2} \operatorname{tr}(\sigma^{AE})^2$$

where U acts on $A = A_1 A_2$.

We note that

$$\| \sigma^{A_2 E} - \sigma^{A_2}_{max} \otimes \sigma^{E} \|_2^2 = \operatorname{tr}(\sigma^{A_2 E})^2 - \frac{1}{|A_2|} \operatorname{tr}(\sigma^{E})^2$$

(because $\operatorname{Tr}(\sigma^{A_2}_{max})^2 = 1/|A_2|$).

Now evaluate

$$\int dU \, \operatorname{tr}(\sigma^{A_2 E}(U))^2$$

$$= \int dU \, \operatorname{tr}_{A_1}(U \sigma^{AE} U^\dagger) \otimes \operatorname{tr}_{A_1'}(U \sigma^{A'E'} U^\dagger) \, S^{A_2 A_2'} \otimes S^{EE'}$$

where $S^{AA'}$ denotes the swap operator on AA'.

Therefore,

$$\int dU \, \operatorname{tr}(\sigma^{A_2 E}(U))^2$$

$$= \operatorname{tr}\left[\sigma^{AE} \otimes \sigma^{A'E'} \left(\int dU \, (U^\dagger \otimes U^\dagger) \, I^{A_1 A_1'} \otimes S^{A_2 A_2'} \, (U \otimes U) \right) \otimes S^{EE'} \right]$$

By the Lemma below, the integral is

$$\int dU \, (U^\dagger \otimes U^\dagger) \, I^{A_1 A_1'} \otimes S^{A_2 A_2'} \, (U \otimes U)$$

$$= c_I \, I^{AA'} + c_S \, S^{AA'} \quad \text{where}$$

$$c_I = \frac{1}{|A_2|} \left(\frac{1 - 1/|A_1|^2}{1 - 1/|A|^2} \right) \leq \frac{1}{|A_2|}, \quad c_S = \frac{1}{|A_1|} \left(\frac{1 - 1/|A_2|^2}{1 - 1/|A|^2} \right) \leq \frac{1}{|A_1|}.$$

Plugging the value of the integral into the trace:

$$\int dU \, \mathrm{tr}\left(\sigma^{A_2E}(U)\right)^2 \leq \frac{1}{|A_2|}\mathrm{tr}\left(\sigma^{E}\right)^2 + \frac{1}{|A_1|}\mathrm{tr}\left(\sigma^{AE}\right)^2$$

and we conclude

$$\int dU \, \| \sigma^{A_2E}(U) - \sigma^{A_2}_{\max} \otimes \sigma^{E} \|_2^2 \leq \frac{1}{|A_1|}\mathrm{tr}\left(\sigma^{AE}\right)^2.$$

From the Cauchy-Schwarz inequality

$$\|M\|_1^2 \leq d\,\|M\|_2^2 \quad \text{and} \quad \langle \sqrt{f} \rangle^2 \leq \langle f \rangle,$$

we find

$$\left(\int dU \, \| \sigma^{A_2E}(U) - \sigma^{A_2}_{\max} \otimes \sigma^{E} \|_1\right)^2 \leq \frac{|A_2E|}{|A_1|}\mathrm{tr}\left(\sigma^{AE}\right)^2$$

— this is the decoupling inequality.

It remains to prove:

Lemma: $$\int dU \, (U^* \otimes U^*)\, I^{A_1A_1'} \otimes S^{A_2A_2'}\, (U \otimes U) = c_I\, I^{AA'} + c_S\, S^{AA'}$$

Proof: The integral commutes with $V \otimes V$, and therefore by Schur's Lemma is a weighted sum of projectors onto irreducible representations. The irreps are the symmetric and antisymmetric tensors, so that

$$\int dU (U^* \otimes U^*)\, I^{A_1A_1'} \otimes S^{A_2A_2'}\, (U \otimes U) = c_{\mathrm{sym}}\, \Pi^{AA'}_{\mathrm{sym}} + c_{\mathrm{anti}}\, \Pi^{AA'}_{\mathrm{anti}}$$

where $\Pi^{AA'}_{\mathrm{sym}}$ projects onto the subspace symmetric under $A \leftrightarrow A'$ and $\Pi^{AA'}_{\mathrm{anti}}$ projects onto the antisymmetric subspace. To compute c_{sym}, evaluate $\mathrm{tr}(\Pi^{AA'}_{\mathrm{sym}} \cdot)$ of both sides. Using $\Pi^{AA'}_{\mathrm{sym}} = \frac{1}{2}\left(I^{AA'} + S^{AA'}\right)$, we obtain

$$\frac{1}{2}\,\mathrm{tr}\left(I^{A_1A_1'}\otimes S^{A_2A_2'}\right)\left(I^{A_1A_1'}\otimes I^{A_2A_2'}+S^{A_1A_1'}\otimes S^{A_2A_2'}\right)$$

$$=\frac{1}{2}\left[\mathrm{tr}\left(I^{A_1A_1'}\otimes S^{A_2A_2'}\right)+\mathrm{tr}\left(S^{A_1A_1'}\otimes I^{A_2A_2'}\right)\right]$$

$$=\frac{1}{2}\left(|A_1|^2|A_2|+|A_1||A_2|^2\right)=C_{sym}\,\mathrm{tr}\,\pi_{sym}^{AA'}=$$

$$=C_{sym}\,\frac{1}{2}|A|(|A|+1)$$

$$\Rightarrow\quad C_{sym}=\frac{|A_1|+|A_2|}{|A|+1}$$

(Here we used $\mathrm{tr}\,S^{AA'}=\mathrm{tr}\left(2\pi_{sym}^{AA'}-I^{AA'}\right)=|A|(|A|+1)-|A|^2=|A|.$)

Similarly, $\pi_{anti}^{AA'}=\frac{1}{2}\left(I^{AA'}-S^{AA'}\right)$ and $\mathrm{tr}\,\pi_{anti}^{AA'}=\frac{1}{2}|A|(|A|-1)$

$$\Rightarrow\quad C_{anti}=\frac{|A_1|-|A_2|}{|A|-1}.$$

Then noting that $C_I=\frac{1}{2}\left(C_{sym}+C_{anti}\right)$

$$C_S=\frac{1}{2}\left(C_{sym}-C_{anti}\right)$$

we obtain

$$C_I=\frac{|A|\cdot|A_1|-|A_2|}{|A|^2-1}=\frac{1}{|A_2|}\,\frac{|A_2|\left(|A_1|^2-1\right)}{|A_2|\left(|A_1|^2-1/|A_2|^2\right)},$$

$$C_S=\frac{|A|\cdot|A_2|-|A_1|}{|A|^2-1}=\frac{1}{|A_1|}\,\frac{|A_1|\left(|A_2|^2-1\right)}{|A_1|\left(|A_2|^2-1/|A_1|^2\right)},$$

which proves the lemma.

Appendix B: Fidelity and L^1 distance

We wish to show:

$$\sqrt{F(\rho,\sigma)} \equiv \| \sqrt{\rho}\sqrt{\sigma} \|_1 = \mathrm{tr}\sqrt{\rho^{\frac{1}{2}}\sigma\rho^{\frac{1}{2}}} \geq 1 - \tfrac{1}{2}\|\rho-\sigma\|_1 .$$

From the polar decomposition of M we obtain

$$\mathrm{tr}\sqrt{M^\dagger M} \geq \mathrm{tr}\, M \Rightarrow \sqrt{F(\rho,\sigma)} \geq \mathrm{tr}(\sqrt{\rho}\sqrt{\sigma}).$$

And

$$\|\sqrt{\rho}-\sqrt{\sigma}\|_2^2 = \mathrm{tr}(\sqrt{\rho}-\sqrt{\sigma})^2 = 2 - 2\,\mathrm{tr}(\sqrt{\rho}\sqrt{\sigma}) \geq 2 - 2\sqrt{F(\rho,\sigma)}$$

$$\Rightarrow \quad \sqrt{F(\rho,\sigma)} \geq 1 - \tfrac{1}{2}\|\sqrt{\rho}-\sqrt{\sigma}\|_2^2$$

Therefore, it suffices to show $\|\rho-\sigma\|_1 \geq \|\sqrt{\rho}-\sqrt{\sigma}\|_2^2$.

Note that $\rho-\sigma = \frac{1}{2}(\sqrt{\rho}-\sqrt{\sigma})(\sqrt{\rho}+\sqrt{\sigma}) + \frac{1}{2}(\sqrt{\rho}+\sqrt{\sigma})(\sqrt{\rho}-\sqrt{\sigma})$,

and we may write $\sqrt{\rho}-\sqrt{\sigma} = \sum_i \lambda_i |i\rangle\langle i| \Rightarrow$

$$|\sqrt{\rho}-\sqrt{\sigma}| = \sum_i |\lambda_i|\,|i\rangle\langle i| = U(\sqrt{\rho}-\sqrt{\sigma}) = (\sqrt{\rho}-\sqrt{\sigma})U$$

where $\{|i\rangle\}$ is the ON basis that diagonalizes $\sqrt{\rho}-\sqrt{\sigma}$

and U is the unitary transformation $U = \sum_i \mathrm{sign}(\lambda_i)\,|i\rangle\langle i|$.

Now, $\mathrm{tr}\,|\rho-\sigma| \geq \mathrm{tr}(\rho-\sigma)U$ (true for any unitary U)

$$= \mathrm{tr}\,|\sqrt{\rho}-\sqrt{\sigma}|(\sqrt{\rho}+\sqrt{\sigma}) = \sum_i |\lambda_i|\langle i|\sqrt{\rho}+\sqrt{\sigma}|i\rangle$$

$$\geq \mathrm{tr}\sum_i |\lambda_i|\langle i|\sqrt{\rho}-\sqrt{\sigma}|i\rangle = \sum_i |\lambda_i|^2 = \|\sqrt{\rho}-\sqrt{\sigma}\|_2^2$$

Thus $\|\rho-\sigma\|_1 \geq \|\sqrt{\rho}-\sqrt{\sigma}\|_2^2$, as we wanted to show.

By the way, it is sometimes convenient to have an upper bound on $F(\rho,\sigma)$ expressed in terms of the L^1 distance $\|\rho-\sigma\|_1$; for example,

$$F(\rho,\sigma) \le 1 - \tfrac{1}{4}\|\rho-\sigma\|_1^2.$$

① First show this is an equality for pure states

$$|\psi\rangle = \begin{pmatrix}\cos\alpha\\ \sin\alpha\end{pmatrix} \quad |\phi\rangle = \begin{pmatrix}\sin\alpha\\ \cos\alpha\end{pmatrix} \Rightarrow |\langle\phi|\psi\rangle|^2 = \sin^2 2\alpha = F(\phi,\psi)$$

$$|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| = \begin{pmatrix}\cos 2\alpha & 0\\ 0 & -\cos 2\alpha\end{pmatrix} \Rightarrow$$

$$\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \|_1^2 = 4\cos^2 2\alpha = 4\left[1 - F(\phi,\psi)\right]$$

② Next note that L^1 distance is monotonic:

$$\|\rho_{AB} - \sigma_{AB}\|_1 \ge \|\rho_A - \sigma_A\|_1.$$

This is true because L_1 distance is optimal distance between prob. distributions for POVM outcomes, and we can perform a POVM on AB that acts nontrivially only on A.

③ Finally, by Uhlmann's theorem,

$$F(\rho_A,\sigma_A) = F(\rho_{AB},\sigma_{AB})$$

where ρ_{AB}, σ_{AB} are purifications with maximal fidelity.

$$= 1 - \tfrac{1}{4}\|\rho_{AB} - \sigma_{AB}\|_1^2$$

$$\le 1 - \tfrac{1}{4}\|\rho_A - \sigma_A\|_1^2,$$

where the last step uses monotonicity of L^1 distance.