Physics 239/139: Quantum information is physical
Spring 2018

Lecturer: McGreevy

These lecture notes live here. Please email corrections to mcgreevy at physics dot ucsd dot edu.

THESE NOTES ARE SUPERSEDED BY THE NOTES HERE

Schrödinger's cat and Maxwell's demon, together at last.

Last updated: 2021/05/09, 15:30:29

Contents

0.1 Introductory remarks

I begin with some discussion of my goals for this course. This is a special topics course directed at students interested in theoretical physics; this includes high-energy theory and condensed matter theory and atoms-and-optics and maybe some other areas, too. I hope the set {students interested in theoretical physics} includes people who do experiments.

The subject will be ideas from information theory and quantum information theory which can be useful for quantum many body physics. The literature on these subjects is sprawling and most of it is not addressed at me. Information theory in general is a lucrative endeavor which was created basically fully formed by the telephone company, and so is all about 'channels' and 'communication'. And much of the literature on quantum information theory is similarly tendentious and product-driven, if somewhat more far-sighted. That is, many these folks are interested in building a quantum computer. Maybe they have already done so; there is a big financial incentive not to tell anyone.

So far, no one has admitted to building a scalable quantum computer. I am not so impatient for humans to get their greedy hands on a quantum computer. In the short term, it will probably make things worse. Nor am I so very interested in most of the engineering challenges which must be overcome to make one. But I find it very interesting to think about the physics involved in making and using one. In particular, there are some beautiful resonances between questions about computation and quantum computation and ideas about phases of matter.

One example is the connection between the quest for a self-correcting quantum memory (a quantum hard drive that you can put in your closet without keeping it plugged in), and the stability of topological order at finite temperature (phases of matter which cannot be distinguished locally). More prosaically, the magnetic hard drives we all use as digital memory rely on spontaneous symmetry breaking. Another example is the deep connection between computationally hard problems (and in particular attempts to solve them with a quantum adiabatic algorithm), and phenomena associated with the word *glass*.

The most important such connection was made famous by Feynman: quantum many body systems manage to find their groundstates and to time evolve themselves. This is a problem which is hard (sometimes provably, quantifiably so) to simulate using a classical computer. How do they do it? This idea of stealing their methods is part of a scientific program which my friend and collaborator Brian Swingle calls 'learning to think like a quantum computer'.

Some other interesting related subjects about which you might provoke me into saying more or less this quarter: Quantum error correction and topological order. Non-abelian anyons, quantum Hall physics. Labels on topological phases in various dimensions. Decoherence, time evolution of open quantum many body systems. Eigenstate thermalization. Quantum algorithms and algorithms for finding quantum states. Tensor network representations.

In case it isn't obvious, I want to discuss these subjects so I can learn them better. For some of these topics, I understand how they can be (and in many cases have been) useful for condensed matter physics or quantum field theory, and I will try to explain them in that context as much as possible. For others, I only have suspicions about their connections to the physics I usually think about, and we'll have to learn them on their own terms and see if we can build some connections.

A word about prerequisites: Talk to me if you are worried. I hope that this class can be useful to students with a diverse set of scientific backgrounds.

Initial plan:

1. Attempt to convey big picture of why the study of quantum many body physics can benefit from careful thinking about quantum information.
2. Sending information through time and space, in a world of adversity.
3. Memory, erasure and the physicality of information.
4. Distinguishing quantum states (distance measures).
5. Groundstate entanglement area law.
6. Quantum error correction and topological order.
7. Consequences of locality.

This is my initial plan; I am open to input about what we should do.

Sources for these notes (anticipated):

Information theory, Inference, and Learning Algorithms, D. MacKay. (!)

Elements of Information Theory, T. M. Cover and J. A. Thomas. (≡ C&T)

Feynman Lectures on Computation, R. Feynman.

Computation, Physics and Information, M. Mézard, A. Montanari.

Lecture Notes on Quantum Information and Quantum Computing, by J. Preskill.

Quantum Information Theory and Quantum Statistics, by D. Petz.

Quantum Information, S. Barnett.

Renner and Christiandl, notes.

Quantum channels guided tour, M. Wolf.

Quantum Information and Quantum Computation, I. Chuang and M. Nielsen.

Classical and Quantum Computation, A. Kitaev, Shen, Vyalyi.

Quantum Information meets Quantum Matter, B. Zeng et al.

Quantum computing since Democritus, by S. Aaronson.

Quantum processes, systems, and information, by B. Schumacher and D. Westmoreland

Quantum Computing, A Gentle Introduction, by E. Rieffel and W. Polak

0.2 Conventions

$\equiv$ means 'equals by definition'. $A \overset{!}{=} B$ means we are demanding that $A = B$.

The convention that repeated indices are summed is always in effect unless otherwise indicated.

$$\ln \equiv \log_e, \quad \log \equiv \log_2 .$$

I'll denote the binary entropy function by $H_2(p) \equiv -p \log p - (1-p)\log(1-p)$ but will sometimes forget the subscript.

Sight is a valuable commodity. In order not to waste it, I will often denote the Pauli spin operators by

$$\mathbf{X} \equiv \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{Y} \equiv \begin{pmatrix} 0 & -\mathbf{i} \\ \mathbf{i} & 0 \end{pmatrix} \quad \mathbf{Z} \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

(rather than $\sigma^{x,y,z}$).

A useful generalization of the shorthand $\hbar \equiv \frac{h}{2\pi}$ is

$$đk \equiv \frac{\mathrm{d}k}{2\pi}.$$

I will also write $\delta\!\!\!/(q) \equiv (2\pi)^d \delta^d(q)$.

I will try to be consistent about writing Fourier transforms as

$$\int \frac{\mathrm{d}^d k}{(2\pi)^d} e^{ikx} \tilde{f}(k) \equiv \int đ^d k \; e^{ikx} \tilde{f}(k) \equiv f(x).$$

WLOG $\equiv$ without loss of generality.
IFF $\equiv$ if and only if.
RHS $\equiv$ right-hand side. LHS $\equiv$ left-hand side. BHS $\equiv$ both-hand side.
IBP $\equiv$ integration by parts.
$+\mathcal{O}(x^n) \equiv$ plus terms which go like x^n (and higher powers) when x is small.
iid $\equiv$ independent and identically distributed.

We work in units where $\hbar$ is equal to one unless otherwise noted.

Please tell me if you find typos or errors or violations of the rules above.

0.3 Lightning quantum mechanics reminder

Axioms of quantum mechanics (QM) for an isolated system.

1. In any physical theory, we have to describe the state of the system somehow. In QM of an isolated system, the state is a vector[1] in a Hilbert space $\mathcal{H}$. By a Hilbert space I mean a vector space over the complex numbers $\mathbb{C}$, equipped with a positive inner product: $\langle\psi|\phi\rangle \in \mathbb{C}$, and $\| \psi \|^2 \equiv \langle\psi|\psi\rangle \geq 0$ with equality only if $|\psi\rangle = 0$.

2. Special bases of $\mathcal{H}$ are determined by *observables*, which are hermitian operators on $\mathcal{H}$, $\mathbf{A} = \mathbf{A}^\dagger$ (the adjoint can be defined by $(\langle a|\mathbf{A}|b\rangle)^\star = \left\langle b|\mathbf{A}^\dagger|a\right\rangle$). Recall that the eigenvectors of a hermitian operator provide an orthonormal basis for $\mathcal{H}$.

3. Time evolution is determined by a special observable, the Hamiltonian $\mathbf{H}$:

$$\mathbf{i}\hbar\partial_t |\psi(t)\rangle = \mathbf{H} |\psi(t)\rangle .$$

4. When we measure the observable $\mathbf{A}$ in the state $|\psi\rangle$, the outcome is an eigenvalue of $\mathbf{A}$ ($\mathbf{A}|a\rangle = a|a\rangle$), and the outcome a occurs with probability $|\langle a|\psi\rangle|^2$. Afterwards, the state is $|a\rangle$.

A single qbit. An example will help. If the Hilbert space is one-dimensional, there is nothing to say – there is only one state. So the simplest example is when $\dim \mathcal{H} = 2$, which is called a *qbit* (sometimes spelled 'qubit'), or spin-$\frac{1}{2}$. Let's introduce a basis of this space by writing

$$\mathcal{H}_2 = \text{span}_\mathbb{C}\{|0\rangle, |1\rangle\},$$

where by $\text{span}_\mathbb{C}\{...\}$, I mean the vector space formed from arbitrary linear combinations of the list of vectors, with coefficients in $\mathbb{C}$. So an arbitrary state in $\mathcal{H}_2$ is of the form

$$z|0\rangle + w|1\rangle, \quad z, w \in \mathbb{C}.$$

The space of such states is the space of ordered pairs (modulo the equivalence (0.1)):

$$\{(z,w)\}/(z,w) \sim \lambda(z,w), \lambda \in \mathbb{C}^\star,$$

[1] An overall multiplication of all states by a nonzero complex number will not change anything,

$$|\psi\rangle \simeq \lambda|\psi\rangle, \quad \lambda \in \mathbb{C}^\star \equiv \mathbb{C}\setminus\{0\}. \tag{0.1}$$

For this reason, it is sometimes said (as I did in lecture) that a state is a *ray* in Hilbert space. Part of this ambiguity can be fixed by normalizing the states, $\| \psi \| \overset{!}{=} 1$.

$\mathbb{CP}^1$, complex projective space, which is geometrically a two-sphere (as you'll show on the first homework), called the *Bloch sphere*. Let's introduce an operator which is diagonal in the given basis:

$$\mathbf{Z}\left|0\right\rangle = \left|0\right\rangle, \quad \mathbf{Z}\left|1\right\rangle = -\left|1\right\rangle$$

(a linear operator is defined by its action on a basis). Notice that $\mathbf{Z}^2 = 1\!\!1_2$, the identity, and $\mathbf{Z} = \mathbf{Z}^\dagger$. In this basis, its matrix elements are

$$\begin{pmatrix} \langle 0| \\ \langle 1| \end{pmatrix} \mathbf{Z}\left(\left|0\right\rangle, \left|1\right\rangle\right) = \begin{pmatrix} \langle 0|\mathbf{Z}|0\rangle & \langle 1|\mathbf{Z}|0\rangle \\ \langle 0|\mathbf{Z}|1\rangle & \langle 1|\mathbf{Z}|1\rangle \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \sigma^z,$$

the Pauli matrix along z. Let's also define an operator $\mathbf{X}$ which satisfies

$$\mathbf{XZ} = -\mathbf{ZX}. \tag{0.2}$$

Then, given an eigenvector of $\mathbf{Z}$,

$$\mathbf{Z}\left|s\right\rangle = (-1)^s\left|s\right\rangle, \qquad s = 0, 1, \tag{0.3}$$

consider the state $\mathbf{X}\left|s\right\rangle$. It has

$$\mathbf{Z}\left(\mathbf{X}\left|s\right\rangle\right) \overset{(0.2)}{=} -\mathbf{XZ}\left|s\right\rangle \overset{(0.3)}{=} -(-1)^s\left(\mathbf{X}\left|s\right\rangle\right).$$

This shows that $\mathbf{X}\left|s\right\rangle = \left|\bar{s}\right\rangle$ – the eigenvector of $\mathbf{Z}$ with the other eigenvalue. That is, $\mathbf{X}$ flips the spin. Its matrix elements in the $\mathbf{Z}$-basis are

$$\begin{pmatrix} \langle 0| \\ \langle 1| \end{pmatrix} \mathbf{X}\left(\left|0\right\rangle, \left|1\right\rangle\right) = \begin{pmatrix} \langle 0|\mathbf{X}|0\rangle & \langle 1|\mathbf{X}|0\rangle \\ \langle 0|\mathbf{X}|1\rangle & \langle 1|\mathbf{X}|1\rangle \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma^x,$$

the Pauli matrix along x. We can also define $\mathbf{Y} \equiv \mathbf{iXZ}$, whose matrix elements are σ^y. Often it is convenient to conflate the matrix elements $\sigma^{x,y,z}$ and the operators $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, particularly because the latter notation writes the important information so much bigger.

What's the big deal about the Pauli matrices? The general observable on a two state system can be written as

$$\mathbf{A} = a_0 1\!\!1_2 + a_i\sigma^i = a_0 1\!\!1_2 + a_x\mathbf{X} + a_y\mathbf{Y} + a_z\mathbf{Z} = \begin{pmatrix} a_0 + a_3 & a_1 - \mathbf{i}a_2 \\ a_1 + \mathbf{i}a_2 & a_0 - a_3 \end{pmatrix}.$$

We can prove this simply by checking that the final expression parametrizes an arbitrary 2×2 matrix.

Finally, we must explain how to compose quantum systems: given a quantum description of system 1, and system 2, what is the Hilbert space of the combined system (which I'll need for example if I want to let them interact)? To answer this, consider the extreme case where we have qbit 1 on Earth and qbit 2 on the other side of the galaxy. Imagine that they do not influence each other in any way, and they've been prepared independently. Then it's reasonable to demand[2] that the probabilities for outcomes of measurements of observables acting separately on the two systems should *factorize* – $P(1,2) = P(1)P(2)$, that is, the outcomes are uncorrelated. This is what we get if we assume the state of the combined system is of the form

$$|a\rangle_1 \otimes |b\rangle_2 \,, \tag{0.4}$$

where $|a\rangle_1 \in \mathcal{H}_1$ is a state of system 1, and $|b\rangle_2 \in \mathcal{H}_2$ is a state of system 2. The symbol $\otimes$ here is a placeholder. If we define the inner product on such states to also factorize:

$$\left(\langle c|_1 \otimes \langle d|_2\right)\left(|a\rangle_1 \otimes |b\rangle_2\right) \equiv \langle c|a\rangle_1 \langle d|b\rangle_2$$

then $P(1,2) = P(1)P(2)$ follows from the measurement rule.

But now axiom 1 tells us our combined Hilbert space must in particular be a vector space – that we must allow arbitrary linear combinations of states of the form (0.4). This means the combined Hilbert space is

$$\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2 \equiv \text{span}_{\mathbb{C}}\{|a\rangle \otimes |b\rangle \,,\;\; |a\rangle_1 \in \mathcal{H}_1, |b\rangle_2 \in \mathcal{H}_2\},$$

the *tensor product* of the two spaces. It quickly becomes inconvenient to write $|a\rangle \otimes |b\rangle$ over and over, and we abbreviate $|a\rangle \otimes |b\rangle \equiv |a,b\rangle$.

Notice that the tensor product of an N-dim'l $\mathcal{H}_N$ and an M-dim'l $\mathcal{H}_M$ has dimension $N \times M$[3]. The general state in $\mathcal{H}_N \otimes \mathcal{H}_M$ has the form

$$|w\rangle = \sum_{n,m} w_{nm} |nm\rangle \neq \left|v^1\right\rangle_N \otimes \left|v^2\right\rangle_M$$

and is *not* a product of a state in $\mathcal{H}_N$ and a state in $\mathcal{H}_M$ – that is only the case if the matrix $w_{nm} = v^1_n v^2_m$ has rank one. Such a state in $\mathcal{H}_N \otimes \mathcal{H}_M$ is called *unentangled.*

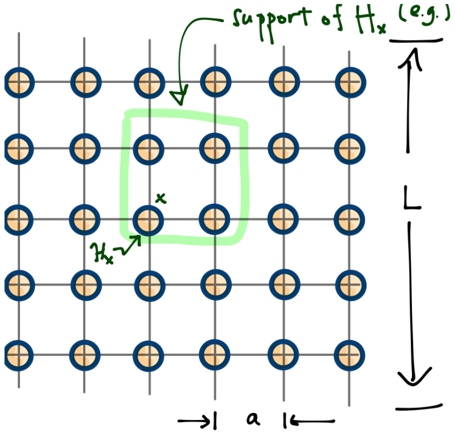
[2] Some folks regard this as a fifth axiom.

[3] This is to be distinguished from the *direct sum* of two spaces,

$$\mathcal{H}_N \oplus \mathcal{H}_M \equiv \text{span}_{\mathbb{C}}\{|n\rangle_N \,, |m\rangle_M\}$$

which has the (generally much smaller) dimension $N + M$. An example where the direct sum obtains is the following: think of $\mathcal{H}_N$ as the Hilbert space of a particle hopping amongst N sites. If I then allow it to hop on M more sites, the resulting Hilbert space is $\mathcal{H}_N \oplus \mathcal{H}_M$. If instead I allow the particle to hop in two dimensions, on an $N \times M$ grid, then the resulting hilbert space is $\mathcal{H}_N \otimes \mathcal{H}_M$.

1 Hilbert space is a myth



In this course we are going to talk about *extensive quantum systems*. A quantum system can be specified by its Hilbert space and its Hamiltonian. By the adjective *extensive* I mean that the Hilbert space is defined by associating finite-dimensional Hilbert spaces $\mathcal{H}_x$ to chunks of *space*, labelled by some coordinates x. Then couple them by a local Hamiltonian, $H = \sum_x H_x$, where H_x acts only on the patch at x and not-too-distant patches (and as the identity operator on the other tensor factors in $\mathcal{H}$).

For example, we can place a two-state system at the sites of a hypercubic lattice. I will call such a two-state system a *qbit* or a *spin*, whose Hilbert space is $\mathcal{H}^{\text{qbit}} \equiv \text{span}_{\mathbb{C}}\{|\uparrow\rangle \equiv |0\rangle\,,\ \ |\downarrow\rangle = |1\rangle\}$.

The phenomena whose study we will find most fulfilling only happen in the *thermodynamic limit*, where the number of patches grows without bound. I will use L to denote the linear size of the system. For a cubic chunk of d-dimensional hypercubic lattice, there are $\left(\frac{L}{a}\right)^d$ patches, where a is the size of the patches. So the thermodynamic limit is $L \to \infty$, or more precisely $L \gg a$. In the mysterious first sentence of this paragraph, I am referring to *emergent* phenomena: qualitatively new effects which can never be accomplished by small systems, such as spontaneous symmetry breaking (magnetism, superconductivity, the rigidity of solids), phase transitions, topological order, and all the other things we have not thought of yet because we are not very smart.[4] [5]

I am making a big deal about the thermodynamic limit here. Let me pause to explain, for example, why no there's no spontaneous symmetry breaking (SSB) in finite volume, classically and quantum mechanically.

[4] In case you doubt that characterization, ask yourself this: How many of the items on this list were discovered theoretically before they were found to occur in Earth rocks by our friends who engage in experiments? The answer is **none**. Not one of them! Let us be humble. On the other hand: this is a source of hope for more interesting physics, in that the set of Earth rocks which have been studied carefully is a very small sample of the possible emergent quantum systems.

[5] Can you think of other elements I should add to this list? One possibility (thanks to Ibou Bah for reminding me) can be called *gravitational order* – the emergence of dynamical space (or spacetime) (and hence gravity) from such emergent quantum systems. The best-understood example of this is AdS/CFT, and was discovered using string theory. I was tempted to claim this as a victory for theorists, but then I remembered that we discovered gravity experimentally quite a while ago.

In a classical system, suppose that our Hamiltonian is invariant under (for definiteness) a $\mathbb{Z}_2$ symmetry: $H(s) = H(-s)$. Then, in equilibrium at coolness β, the magnetization is

$$\langle s \rangle \propto \sum_s e^{-\beta H(s)} s = \sum_{\tilde{s} \equiv -s} e^{-\beta H(-\tilde{s})} (-\tilde{s}) = \sum_{\tilde{s} \equiv -s} e^{-\beta H(\tilde{s})} (-\tilde{s}) \propto - \langle s \rangle$$

and hence it vanishes. The remarkable thing is that SSB can happen (in the thermodynamic limit).

The same is true quantumly. A stationary state (including the groundstate) of a system with a finite dimensional Hilbert space cannot break a(n Abelian) symmetry of its Hamiltonian.

Suppose we have a $\mathbb{Z}_2$ symmetry represented by the operator g, $g^2 = 1$. $[g, H] = 0$. A stationary state satisfies $H |\psi\rangle = E |\psi\rangle$, and it is not symmetric if $g |\psi\rangle = |\psi_g\rangle \neq |\psi\rangle$. This implies $|\psi_g\rangle$ is also an eigenstate with the same energy. But now what's to stop us from adding g to the Hamiltonian, $H \mapsto H + g$?[6][7] If H contains such a term, then there is tunneling between $|\psi\rangle$ and $|\psi_g\rangle$ and neither is stationary; only the uniform-magnitude linear combinations (eigenstates of g) are eigenstates of H, with distinct eigenvalues. The dramatic phenomenon is that the tunneling rate can depend on L (because the symmetry generator g itself is *not* a local operator, and can only be made by multiplying together many terms from the Hamiltonian), so that the overlap between the different groundstates goes to zero in the thermodynamic limit.

This statement plays a starring role in the *More is Different* paper. In that regard, it is worth noting that SSB is a class of emergent phenomena, not the only one, and as I describe next, not a very quantum mechanical one.

[6] Possible smarty-pants answer: non-Abelian symmetry. If the group is non-Abelian, we can't add any of the generators to H preserving the whole group. An example is the $\mathsf{SU}(2)$ ferromagnet. This really does have a degenerate set of groundstates in finite volume without fine-tuning. The better definition of SSB which excludes this requires reference to the response to an external symmetry-breaking field, and specifically, whether :

$$\partial_h f(h)|_{h\to 0^+} \overset{?}{=} \partial_h f(h)|_{h \to 0^-}$$

(Here I'm describing a classical system and f is the free energy; for a quantum system, we should use the groundstate energy instead.) This discontinuity in the magnetization requires a singularity in the function $f(h)$, which can only happen in the thermodynamic limit. A good, brief definition of SSB (which incorporates all of these subtleties and rules out the finite-size ferromagnet) is that it is associated with a diverging susceptibility $\partial_h^2 f|_{h=0}$, where diverging means 'diverging in the thermodynamic limit'. So $L \to \infty$ is built in. (Thanks to Wang Yang for asking me about the finite-size ferromagnet.)

[7] Here I am building in the theoretical prejudice that a good model of the system should be *generic*, that is, its physics should remain valid in an open set in the space of Hamiltonians consistent with the symmetries around the model Hamiltonian of interest.

So maybe now you believe that it matters to take $L/a \gg 1$. The whole Hilbert space of our extensive quantum system is then

$$\mathcal{H} = \otimes_x^{\mathcal{N}} \mathcal{H}_x \ ,$$

where I've used $\mathcal{N} \equiv \left(\frac{L}{a}\right)^d$ to denote the number of patches.

Suppose that a basis of the local Hilbert space $\mathcal{H}_x$ is $\{|s_x\rangle \, , s_x = 1..\mathfrak{D}\}$, so that the general state in this space can be labelled as

$$\mathcal{H}_x \ni \sum_{s_x=1..\mathfrak{D}} c_{s_x} |s_x\rangle$$

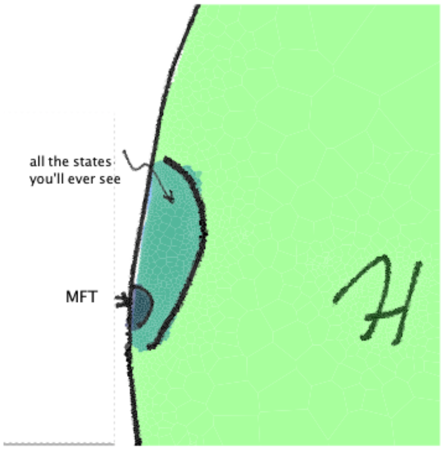
with $\mathfrak{D}$ complex numbers c_{s_x}. (You can take $\mathfrak{D} = 2$ if you insist on qbits.)

By definition of the tensor product, the general state in the full $\mathcal{H}$ is then of the form

$$|\psi\rangle = \sum_{\{s_x=1..\mathfrak{D}\}} c_{s_1 ... s_{\mathfrak{D}\mathcal{N}}} |s_1 ... s_{2\mathcal{N}}\rangle \ . \tag{1.1}$$

That is, we can represent it as a vector of $\mathfrak{D}^{\mathcal{N}}$ complex numbers, $c_{s_1 ... s_{\mathfrak{D}\mathcal{N}}}$.

Everything I've said so far, characterizing quantum systems in terms of their Hilbert spaces, is true. But there are several very serious problems with this description of a quantum many body system. The first and most immediate is that this is too many numbers for our weaks and tiny brains. **Exercise:** Find the number of qbits the dimension of whose Hilbert space is the number of atoms in the Earth. (It's not very many.) Now imagining diagonalizing a Hamiltonian acting on this space.



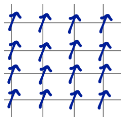
The other reasons for the title of this section are not quite so easy to explain, and part of our job this quarter is to explain them. The basic further statement is: you can't get there from here. Most states in $\mathcal{H}$ cannot be reached by time evolution with any local Hamiltonian for any finite time, starting with a product state. (Why am I assuming 'here' is a product state? More below.) For more rhetoric along these lines, I recommend *e.g.* this discussion. I'll say more about this result in §1.3.

How is it that there is a thriving theory of condensed matter physics which does have something to say about the list of fulfilling emergent phenomena I described above, which *only* happen when the dimension of the Hilbert space is so ginormous?? (How could anyone possibly think we have understood all there is to understand about this?)

One reason there is such a thriving theory is that *ground states of local Hamiltonians are special.* There has been a lot of progress on understanding how they are special in the past X years, a slogan for which is *the Area Law for Entanglement.* Groundstates are less entangled than the vast majority of states of the form (1.1). To start giving meaning to these words, let me start by saying that this means that they are on the same planet as mean field theory:

1.1 Mean field theory is product states

Mean field theory means restricting attention to states of the form



$$|\psi_{\rm MF}\rangle = \otimes_x \left(\sum_{s_x=1..\mathfrak{D}} c_{s_x} |s_x\rangle \right) . \tag{1.2}$$

States which can be factorized in this way (in some factorization of $\mathcal{H}$) are called *unentangled* (with respect to that factorization of $\mathcal{H}$). This writes the state in terms of only $\mathcal{N}\mathfrak{D}$ numbers c_{s_x}, a vast reduction.

The name 'mean field theory' connotes the idea (commonly applied *e.g.* to models of classical magnets) of considering the experience of a single spin, and treating the effects its neighbors as a single field (the mean field). It is possible to derive (see *e.g.* here (section 4)) this usual mean field theory of classical magnets by a variational ansatz for the probability distribution which is *factorized*: $p(s) = \prod_x p(s_x)$. That is: the free energy computed with this distribution gives a variational bound on the correct equilibrium Boltzmann distribution free energy. In the same spirit, think of the expression (1.2) as a variational ansatz with $\mathcal{N}\mathfrak{D}$ variational parameters.

An example: the transverse field Ising model (TFIM). A previous time I taught a special topics course, I spent most of it talking about this model, because there's so much to say about it, and I promised myself I wouldn't do that again. Nevertheless...

Place qbits at the sites of some graph. Let

$$H_{\rm TFIM} = -J \left(\sum_{\langle ij \rangle} Z_i Z_j + g \sum_i X_i \right) .$$

Here $\langle ij \rangle$ indicates the the site i and j share a link. The first term is a ferromagnetic (if $J > 0$) interaction between neighboring spins, diagonal in the Z-basis. The name of the model comes from the fact that the term gJX_i is a Zeeman energy associated with a magnetic field in the x direction, transverse to the direction in which the ferromagnetic term is diagonal. These terms don't commute with each other.

When $g = 0$, it's easy to find groundstates: just make all the spins agree:

$$|\Uparrow\rangle \equiv |\uparrow\uparrow\uparrow \ldots\rangle\,, \quad |\Downarrow\rangle \equiv |\downarrow\downarrow\downarrow \ldots\rangle \tag{1.3}$$

are exact groundstates, in which the spins are unentangled. However, the states

$\Big|$ $_\pm \Big\rangle \equiv \frac{1}{\sqrt{2}} \left(|\Uparrow\rangle \pm |\Downarrow\rangle \right)$

are also groundstates of $H_{g=0}$, and they are entangled. When g is a finite nonzero value, the true groundstate is not a product state. At $g = \infty$ we can ignore the ZZ term and the groundstate is again a product state:

$$|\Rightarrow\rangle = \otimes_x |\rightarrow\rangle_x = \otimes_x \left(\frac{|\uparrow\rangle + |\downarrow\rangle}{\sqrt{2}} \right). \tag{1.4}$$

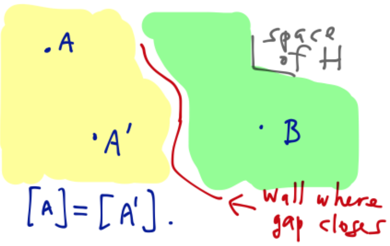
On the homework you'll get to find the best mean field state at various g.

[End of Lecture 1]

Why does mean field theory work, when it does? This depends on what we mean by 'work'. If we mean do a good job of quantitatively modeling the phenomenology of Earth rocks, then that's a difficult question for another day. A more basic and essential goal for our candidate groundstate wavefunction is that it represents the right *phase of matter* (as the true groundstate of H, or as the true groundstate of the true H, since H is only a model after all).

Digression on equivalence classes of gapped systems (please see the beginning of the Spring 2014 239a notes for more discussion of this):

For systems with an energy gap (the first excited state has an energy which is bigger than the groundstate energy by an amount which stays finite when $L \to \infty$), we can make a very sharp definition of what is a phase: all the states that can be reached by continuously deforming the Hamiltonian without closing the energy gap are in the same phase.



Given two gapped Hamiltonians, how can we know whether there is a wall of gaplessness separating them? One way to know is if they differ by some *topological quantity* – something which cannot change continuously, for example because it must be an integer. An example is the number of groundstates: if a system spontaneously breaks a $\mathbb{Z}_2$ symmetry, it must have two groundstates related by the symmetry. If it has a symmetric groundstate, then there is only one. The TFIM has two phases which can be distinguished in just this way (the ferromagnetic (symmetry-broken) phase at $g < 1$ where there are two groundstates and the paramagnetic phase at $g > 1$ where there is a unique symmetric groundstate) .

In the case of the TFIM, mean field theory actually works really well, and that's because *both* phases have representatives whose groundstates which are product states, namely $g = \infty$, where the groundstate is (1.4) and $g = 0$ where the groundstates are (1.3).

Mean field theory is great and useful, and is responsible for much of our (meagre) understanding of quantum many body physics. It does a good job of illustrating SSB. But it is too far in the other direction from (1.1). There is more in the world! One example, which we know exists both platonically and in Earth rocks (at least it can be made to happen in Earth rocks with some encouragement in the form of big magnetic fields and high-quality refrigeration), is *topological order*. This is a phase where there is *no* product-state representative. Another way to say what topological order is: Two phases can be distinct, but have all the same symmetry properties (for example: no symmetries). Another symptom is *long-range entanglement*. I'm going to say much more about this.

All of statistical physics and condensed matter physics is evidence that qualitatively new things can happen with large numbers. So the absolute intractability of many body Hilbert space is an opportunity.

1.2 The local density matrix is our friend

A useful point of view about mean field theory is the 'molecular field' idea: we imagine the experience of a subset A of the system (at its most extreme, a single spin). The rest of the system $\bar{A}$ then behaves as an environment for the subsystem of interest. But in extensive, motivic systems (meaning H is determined by a pattern that repeats itself over different regions of space), we can expect each such subset to have the same experience, and this expectation can be used to derive a set of self-consistent equations.

In a classical stat mech model, the environment determines the local field. In the absence of correlations between the spins, we can do the sum over a single spin without worrying about the others. (I refer to the discussion in these notes for more on the classical case.) Quantum mechanically, there is a new obstacle, beyond mere correlations. This is *entanglement* between a subsystem and the rest of the system.

It's a bit unfortunate that the name for this is a regular word, because it makes it seem imprecise. Given a state $|\psi\rangle \in \mathcal{H}$, and a choice of factorization $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, the two subsystems A and B are *entangled* in the state $|\psi\rangle$ if $|\psi\rangle$ is not a product state, *i.e.* does not factorize in the form $|\psi\rangle \overset{?}{=} |a\rangle_A \otimes |b\rangle_B$.

This new ingredient is a big deal for the subsystem A whose experience we are channeling: if the groundstate of H is entangled between A and $\bar{A}$, it means that A *does not have a wavefunction* of its own! That is: in this case, unless we also measure something in $\bar{A}$, we are uncertain about the wavefunction of A.

This is a very important point, which is the essence of quantum mechanics (never mind those silly superposition tricks, which you can do with ordinary classical light), so let me be very explicit.

A general state

$$|w\rangle = \sum_{i,m} w_{im} |i\rangle_A \otimes |m\rangle_B \neq \left|v^A\right\rangle_A \otimes \left|v^B\right\rangle_B$$

for any $v^{A,B}$. This is only possible if the coefficient matrix factorizes as $w_{i,m} \overset{?}{=} v_i^A v_m^B$. A matrix that can be written this way has rank 1 – only a one-dimensional eigenspace of nonzero eigenvalues.

A crucial point: if we only have access to the stuff in A, then all the operators we can measure have the form $\mathbf{M} = \mathbf{M}_A \otimes \mathbb{1}_{\bar{A}\equiv B}$ – they act as the identity on the complement of A. In any state $|w\rangle$ of the whole system, the expectation value of any such operator can be computed using only the *reduced density matrix* $\boldsymbol{\rho}_A \equiv \mathrm{tr}_{\bar{A}} |w\rangle \langle w|$.[8] This operation by which we obtained $\boldsymbol{\rho}_A$ is called *partial trace*.

The density matrix $\boldsymbol{\rho}_A$ is a positive (and hence Hermitian) operator with unit trace.[9] These are general conditions on any density matrix which allow for a probability

[8] Explicitly,

$$\begin{aligned}\langle \mathbf{M}_A \rangle &= \langle w| \mathbf{M}_A \otimes \mathbb{1}_B |w\rangle = \sum_{j,s} \sum_{i,r} w_{js}^\star \langle j|_A \otimes \langle s|_B \left(\mathbf{M}_A \otimes \mathbb{1}_B\right) w_{ir} |i\rangle_A \otimes |r\rangle_B \\ &= \sum_{ij,r} w_{ir} w_{jr}^\star \langle j|_A \mathbf{M}_A |i\rangle_A = \mathrm{tr}_A \boldsymbol{\rho}_A \mathbf{M}_A,\end{aligned}$$

with

$$\boldsymbol{\rho}_A = \mathrm{tr}_{\bar{A}} |w\rangle \langle w| = \sum_{ij,r} |i\rangle_A {}_A\langle j| \, w_{ir} w_{jr}^\star \ , \quad (\boldsymbol{\rho}_A)_{ij} = \sum_r w_{ir} w_{jr}^\star. \tag{1.5}$$

In (1.5) I assumed that the basis $\{|r\rangle_B\}$ was orthonormal, so that $\langle s|r\rangle = \delta_{s,r}$.

[9] A positive operator $\mathbf{A}$ is one for which $\langle b| \mathbf{A} |b\rangle \geq 0$ for all states $|b\rangle$. Beware that one may encounter an alternative definition that all the singular values (s such that $\det(s\mathbb{1} - \mathbf{A}) = 0$) are positive. These differ for operators with Jordan blocks, like $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ which are positive by the latter definition but not the first. Thanks to Sami Ortoleva for the warning.

interpretation of expectation values $\langle \mathbf{M}_A \rangle = \mathrm{tr}_A \boldsymbol{\rho}_A \mathbf{M}_A$, and here they follow from the normalizedness of the state $|w\rangle$. As with any hermitian matrix, $\boldsymbol{\rho}_A$ can be diagonalized and has a spectral decomposition:

$$\boldsymbol{\rho}_A = \sum_\alpha p_\alpha |\alpha\rangle \langle\alpha|$$

with $\mathrm{tr}_A \boldsymbol{\rho}_A = \sum_\alpha p_\alpha = 1$. $p_\alpha \in [0,1]$ can be regarded as the probability that the subsystem is in the state $|\alpha\rangle$.

The rank of the matrix w is called the *Schmidt number* of the state $|w\rangle$; $|w\rangle$ is entangled if the Schmidt number is bigger than 1. The Schmidt number is therefore also the rank of the reduced density matrix of A. When the Schmidt number is one, the one nonzero eigenvalue must be 1, so in that case the density matrix is a projector onto a pure state of the subsystem.

Entanglement is not the same as correlation (though there is a correlation). These two spins are (perfectly) correlated:

$$|\uparrow\rangle \otimes |\uparrow\rangle$$

but not (at all) entangled: they do actually have their own wavefunctions.

So the Schmidt rank is one way to quantify (by a single number) how entangled A and its complement are in the state $|w\rangle$. Since I will use it all the time, I might as well mention now that an often-more-useful measure is the *von Neumann entropy* of $\boldsymbol{\rho}_A$:

$$S[\boldsymbol{\rho}_A] \equiv -\mathrm{tr}_A \boldsymbol{\rho}_A \log \boldsymbol{\rho}_A .$$

So: really the right local question to ask, to extend mean field theory beyond product states, is: what is the reduced density matrix of our subsystem, A, when the whole system is in its groundstate, and what is its experience of the world.

I want to advocate the following analogy, to motivate the plan of our course this quarter: think of our heroic little subsystem A as a quantum computer. It is a quantum system, perhaps coherent, trying to quantumly compute (for example) its own groundstate. (Does it do this by writing it out as a vector of $\mathfrak{D}^{|A|}$ complex numbers and doing row-reduction? Probably not.) But it is subject to a noisy environment, in the form of the rest of the system. What is *noise*? In its usage in science (and often colloquially too) it is something that we're not paying enough attention to, so that we are unable to resolve or keep track of its details. The rest of the system keeps interacting with our poor subsystem, trying to measure its state, decohering[10] it. Some

[10] Do you want me to say more about this? Here's the short version:

$$\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)_A \otimes |0\rangle_E \overset{\text{wait}}{\rightsquigarrow} \frac{|00\rangle + |11\rangle}{\sqrt{2}} \overset{\text{ignore } E}{\rightsquigarrow} \boldsymbol{\rho}_A = \frac{1}{2}\mathbb{1}.$$

local rules (H_x) for the subsystem's behavior will do better than others at this. These are just the kinds of things that people have to worry about when they are engineering (or imagining someday telling someone how to engineer) a quantum computer.

So, partly motivated by this analogy, we are going to try to understand what is known about *open* quantum systems, quantum systems subject to some environment, which we may model at various levels of detail.

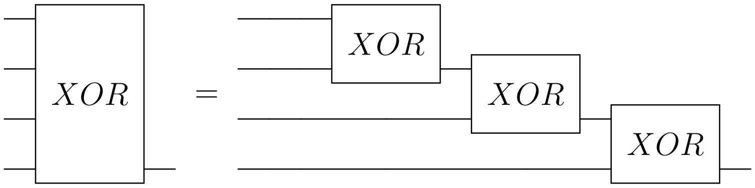
For better or worse, quite a bit is known about this subject, some of it quite rigorously so. And most of it builds on analogous results regarding the communication and storage of classical information. So we're going to spend some time on that.

1.3 Complexity and the convenient illusion of Hilbert space

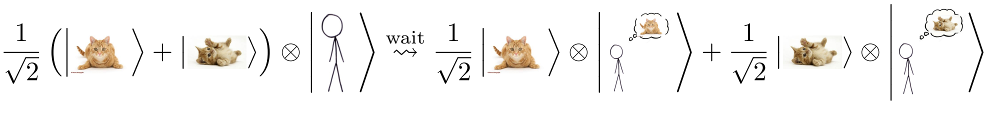
But first: Since it will give me an opportunity to illustrate a nice resonance between the theory of computation (specifically a result of Shannon) and quantum many body physics, I will say more precisely what is the statement of 'you can't get there from here'.

Classical circuit complexity. First, consider the set of Boolean functions on n bits, $f : \{0,1\}^n \to \{0,1\}$. How many of these are there? We have to specify what the function does to every configuration of the input bits, and there are two choices for each, so there are 2^{2^n} such functions. That grows rapidly with n, just like the dimension of many-body Hilbert space $\dim \mathcal{H}$.

Suppose we want to make computers to compute such functions (with large n), by building them out of some set of elementary 'gates' – functions which act on just a few bits at a time. For example, we can build the XOR on n bits (which adds the bits mod two[11] out of $n-1$ successive pairwise XORs:



Here's a slightly more detailed version of that first step:



(You should recognize the observer depicted here from xkcd. The cat pictures are of unknown provenance.)

[11] Thanks to Ahmed Akhtar and Aneesh Manohar for pointing out my error.

In this circuit diagram, time is running to the right (sorry). The lines coming in from the left represent the n input bits, and the one line coming out is the outbit bit. A circuit diagram is a kind of Feynman diagram – a diagram which associates a number with a physical process. (I'll say more about this.)

One way to measure the *complexity* of a function f is by the minimum number of 2-bit gates needed to compute it. By changing the elementary gates you might be able to change the answer a bit. One well-tested, universal, sharp distinction is how that number of gates scales with n. In particular, whether it is polynomial in n or exponential in n (or something else) can't be changed by changing the list of elementary gates. (As usual, 'universal' means independent of short-distance details.)

(Another measure of complexity we might consider is the (minimum) *depth* of the circuit, which is the maximum number of gates a bit needs to traverse to get from input to output.)

Are all boolean functions computable with a number of gates that grows like a polynomial in the input size n? Shannon answered this question with a counting argument: First count how many circuits we can make with n inputs and T k-input gates. Each such circuit computes one function (some circuits may compute the same function, so this is a lower bound). For each gate we have $n+T$ choices for each input, so there are $((n+T)^k)^T$ such circuits. We need

$$(n+T)^{kT} \geq 2^{2^n} \tag{1.6}$$

to compute all the functions, so we require

$$kT\log(n+T) \geq 2^n, \quad \implies \quad T \geq \frac{2^n}{k\log(n+T)} \geq \frac{2^n}{kn}.$$

We conclude that for *most* functions, the number of required gates grows exponentially in n. Allowing for m types of elementary gates doesn't help: it changes the number of circuits to just $(m(n+T)^k)^T$.

Unfortunately this argument is not constructive and most functions that you can actually describe concretely and easily will be computable with poly(n) gates. Maybe you want an example of one that can't. It was apparently a big deal when one was found (by Hartmanis and Stearns in 1965), building on Turing's demonstration of the existence of functions which aren't computable at all. I refer you to Scott Aaronson's notes for this, but briefly: The hard problem in question asks whether a Turing machine halts after $f(n)$ steps (for example you could take $f(n) = e^{an}$ for any a). This problem takes any Turing machine at least $f(n)$ steps to solve. If not you can make a contradiction as follows: Given a machine which solves the problem faster than $f(n)$, use it to build a machine P which takes a Turing machine M as input and (a) runs forever if M halts before $f(n)$ or (b) halts if M runs for longer than $f(n)$ steps. So if P doesn't halt by $f(n)$ it never will. Now feed P to itself. Then we rely on the equivalence of computational models, that is, anything you can do efficiently with a Turing machine can be simulated with a circuit.

Quantum circuits. The result of Poulin *et al.* is basically a quantum version of Shannon's result. Instead of functions on n bits, consider the Hilbert space

$$\mathcal{H} = \otimes_{i=1}^{n} \mathcal{H}_i$$

where I will assume WLOG that $\mathcal{H}_i$ is a qbit (if it's not, break it into more factors and if necessary throw some away at the end). We'll consider a Hamiltonian

$$H = \sum_{X \subset \{1...n\}} H_X(t)$$

where $H_X(t)$ acts only on the subset X, and can depend arbitrarily on time, and the subsets need have no notion of locality. But: we assume that the support of each term H_X is $|X| \leq k \sim n^0$ – finite in the thermodynamic limit $n \to \infty$. (Such a Hamiltonian is called k-local; a local Hamiltonian is a special case.)

The question they ask is: which states can we reach (say, starting from a product state) by time evolution with such a k-local Hamiltonian for a time which is polynomial in the system size, $t \sim n^\alpha$? The answer is not very many of them.

Time evolution. Recall the QM axiom for time evolution:

$$\mathbf{i}\partial_t |\psi(t)\rangle = H(t) |\psi(t)\rangle \tag{1.7}$$

(we are allowing the Hamiltonian $H(t)$ to depend on time). We can solve this equation by introducing the time evolution operator $U(t)$ such that $|\psi(t)\rangle = U(t) |\psi(0)\rangle$. Then (1.7) is satisfied if

$$\mathbf{i}\partial_t U(t) = H(t)U(t), \tag{1.8}$$

with the initial condition $U(0) = \mathbb{1}$. Here's a solution of (1.8):

$$U(t) = \mathbb{1} - \mathbf{i}\int_0^t dt_1 H(t_1)U(t_1).$$

The only shortcoming of this solution is that it has U again on the RHS. We can do a little better by substituting this equality again for the U on the RHS:

$$U(t) = \mathbb{1} - \mathbf{i}\int_0^t dt_1 H(t_1)\left(\mathbb{1} - \mathbf{i}\int_0^{t_1} dt_2 H(t_2)U(t_2)\right).$$

Perhaps we should keep doing this a few more times:

$$\begin{aligned} U(t) &= \mathbb{1} - \mathbf{i}\int_0^t dt_1 H(t_1) + (-\mathbf{i})^2\mathbf{i}\int_0^t dt_1 \int_0^{t_1} dt_2 H(t_1)H(t_2) \\ &\qquad + (-\mathbf{i})^3\mathbf{i}\int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 H(t_1)H(t_2)H(t_3) + \cdots \qquad (1.9) \\ &= \mathcal{T}\Bigg(\mathbb{1} - \mathbf{i}\int_0^t dt_1 H(t_1) + \frac{1}{2}(-\mathbf{i})^2\mathbf{i}\int_0^t dt_1 \int_0^t dt_2 H(t_1)H(t_2) \\ &\qquad + \frac{1}{3}(-\mathbf{i})^3\mathbf{i}\int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 H(t_1)H(t_2)H(t_3) + \cdots\Bigg) \qquad (1.10) \\ &= \mathcal{T}\sum_{\ell=0}^{\infty}\frac{1}{\ell!}\left(-\mathbf{i}\int_0^t dt' H(t')\right)^\ell = \mathcal{T}e^{-\mathbf{i}\int_0^t dt' H(t')}. \end{aligned}$$

The tricky step here is the introduction of the *time-ordering* operation $\mathcal{T}$: $H(t)$ and $H(t')$ don't commute, so the order matters. In (1.9) they appear from left to right in the order of their arguments: $H(t)$ is to the left of $H(t')$ if $t > t'$. The operation $\mathcal{T}$ puts them in the right order:

$$\mathcal{T}H(t_1)H(t_2) = \theta(t_1 - t_2)H(t_1)H(t_2) + \theta(t_2 - t_1)H(t_2)H(t_1)$$

(where $\theta(t)$ is the Heaviside function). In that same step, we extend the range of integration from the simplex $t_1 \geq t_2 \geq t_3 ... \geq t_\ell \geq 0$ to the cube $t_i \in [0, t]$, which is $\ell!$ times as big.

Their argument has two parts.

(1) Trotterize: The first idea is that the unitary, continuous Hamiltonian time evolution can be approximated arbitrarily well by a quantum circuit made of unitary operators acting on k qbits at a time. The time evolution operator from time 0 to time t is

$$U(t) = \mathcal{T}e^{-\mathbf{i}\int_0^t ds H(s)} \simeq \prod_{p=1}^{N_p} U_p \equiv \prod_p e^{-\mathbf{i}H_{X_p}(t_p)\Delta t_p} \equiv U_{TS}. \qquad (1.11)$$

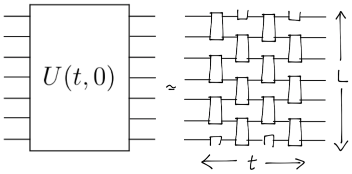
In the product on the RHS, the consequence of the time-ordering is that the factor U_p is to the left of $U_{p'}$ if $t_p > t_{p'}$. This approximation is sometimes called Trotter-Suzuki decomposition and is used in the derivation of the path integral. Errors come from (a) ignoring variation of $H(t)$ on timescales small compared to Δt, which is fine if $\Delta t \ll \parallel \partial_t H \parallel^{-1}$. Here $\parallel \mathcal{O} \parallel \equiv \sup_{\{\text{normalized states, } |\psi\rangle\}} \parallel \mathcal{O}|\psi\rangle \parallel$ is the operator norm. The second source of error is (b) the fact that the terms in H at different times and different X need not commute. The Baker-Campbell-Hausdorff formula can be used to show that

$$\parallel U - U_{TS} \parallel \leq c^2 (\Delta t)^2 L^2$$

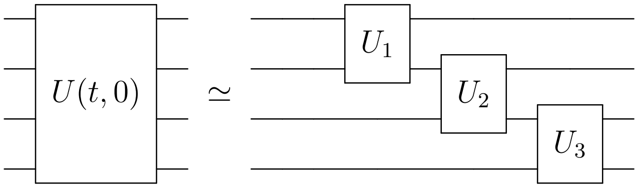
where U_{TS} is the circuit approximation and the constant is $c \sim \max_{X_1,X_2} \| [H_{X_1}, H_{X_2}] \|$. If we demand a total error ϵ in our circuit approximation to the time evolution, and there are L terms in the Hamiltonian (L grows with n) then the number of gates we need is

$$N_p = L\frac{t}{\Delta t} = \frac{c}{\sqrt{\epsilon}} t^{3/2} L^2,$$

the important point being that this is a polynomial in t and L (though I'm finding a different power than the paper by Poulin *et al*). Here, by our assumption about H_X, U_p is a ($\leq k$)-body unitary operator – it acts on only k of the n qbits. The figure at right illustrates $k = 2$.



Furthermore, the factors in (1.11) are time-ordered, $t_p \geq t_{p-1}$. So the circuit might look something like this, for $k = 2$ (and $n = 4$):

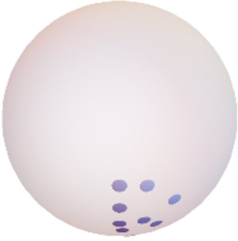


(2) Count balls. Now let's ask which states can be made by such Hamiltonians in a time polynomial in n, starting with some reference state. The assumption on t implies that the number of k-qbit gates needed to approximate $U(t,0)$ goes like n^α for some α. The number of circuits we can make from these is (just as in the classical case (1.6))

$$N_{\text{circuits}} \sim \left(mn^{2k}\right)^{n^\alpha}$$

where m is the number of gate types, and n^k is the number of subsets of degrees of freedom on which each k-qbit gate can be applied. As in the classical case, N_{circuits} bounds from above the number of distinct states we can make.

[End of Lecture 2]



Let's allow an error ϵ, so we declare victory if we get inside a ball of radius ϵ from the desired state. The volume of the $((D \equiv 2 \cdot 2^n - 1)$-real-dimensional) ball around the output of each circuit is

$$V_\epsilon = \epsilon^D \frac{\pi^{D/2}}{\Gamma\left(\frac{D+2}{2}\right)} \simeq \epsilon^{2 \cdot 2^n} \frac{\pi^{2^n}}{\Gamma\left(2^n\right)}.$$

The normalized states in $\mathcal{H}$ live on a unit sphere with $2 \cdot 2^n - 1$ real dimensions; its volume is

$$S_{\mathcal{H}} = \frac{2\pi^{2^n}}{\Gamma\left(2^n\right)}.$$

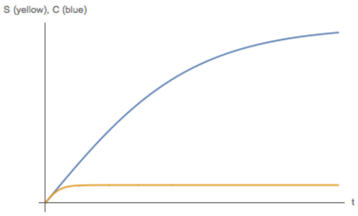
What fraction of this do we cover with our poly-n circuits? Only

$$f = \frac{N_{\text{circuits}} V_\epsilon}{S_{\mathcal{H}}} \sim \epsilon^{2^n} n^{n^\alpha} \overset{n \to \infty, \epsilon < 1}{\to} 0,$$

a doubly-exponentially tiny fraction. It's the powers of ϵ that get us.

So this is what I meant by 'you can't get there from here' – time evolution by a local hamiltonian for an amount of time polynomial in system size covers only a tiny fraction of all states. Note that it's not clear that accessibility by time evolution from product states is the only notion of 'physical'. For example, (Tarun Grover points out that) it could be that excited eigenstates of local Hamiltonians are not accessible in this sense.

How do we distinguish between states we can make and states we can't? We can call it the *complexity*. It will saturate at the time when we can make all the states, and evolving longer just makes the same states again. It is actually *not* the entanglement between the constituents which continues to grow – the entanglement entropy (shown in yellow at right) of a subsystem saturates at $S \sim R$, where R is the size of the subsystem. This can happen in a reasonable amount of time, and actually happens when a system starts in its groundstate, gets kicked and then thermalizes at some finite temperature.



I haven't really defined entropy yet. That's next.

While I'm at it, here is one more reason to say that $\mathcal{H} = \otimes_{i=1}^{N} \mathcal{H}_x$ is an illusion (in the thermodynamic limit). This is that many of the properties of Hilbert space that we hold dear (and which are assumptions in our theorems about it) rely on the property that $\mathcal{H}$ is *separable*. This means that it has a countable basis. If we have a half-infinite ($N \to \infty$) line of qbits and we take seriously the basis

$$\mathcal{H} = \text{span}\{|s_1 s_2 s_3 \cdots\rangle\,,\ s_i = 0 \text{ or } 1\}$$

then the argument of the ket is precisely the binary decimal representation of a real number between 0 and 1. Cantor's diagonal argument shows that this set is not countable.[12] (Propose a countable basis. Then line up the basis elements in a big vertical table. Make a new number by flipping the nth digit of the nth entry in the table. You've made a number not in the list, and hence a state which cannot be made by a linear combination of the others.)

The resolution of this issue is that the Hamiltonian provides extra information: most of the crazy states which are causing the trouble (and making us think about awful real analysis issues) do not have finite energy for any reasonable Hamiltonian.

Postscript to chapter 1: I learned from the lectures of Wolf about this quote from von Neumann:

> *"I would like to make a confession which may seem immoral: I do not believe in Hilbert space anymore."*

[J.von Neumann in a letter to Birkhoff, 1935]

This point of view led to the study of von Neumann algebras and axiomatic quantum field theory. Somehow I still have some hope for it.

[12] I wish I had a useful reference for this discussion. I learned about it from Henry Maxfield, Kenan Diab, and Lauren McGough.

2 Quantifying information

Probability theory is a (weirdly important) subset of quantum mechanics. As E.T. Jaynes says, science is reasoning with incomplete information. Sometimes it is useful to quantify that information. This is the job of probability theory.

I will speak about probability distributions $p_x \equiv p(x)$ on discrete, finite sample sets $x \in \mathcal{X}$, $|\mathcal{X}| < \infty$. The probability interpretation requires $\sum_{x\in\mathcal{X}} p_x = 1$. I will sometimes conflate the random variable X with its values x, as in the ubiquitous but meaningless-if-you-think-about-it-too-much equation

$$\langle x \rangle \equiv \sum_{x\in\mathcal{X}} p_x x.$$

When I want to do a little better I will write things like

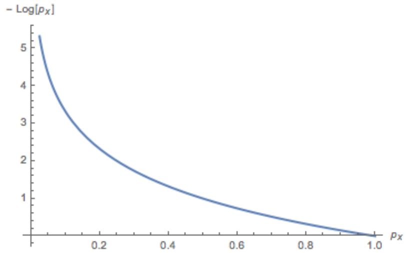
$$\langle X \rangle_X \equiv \sum_{x\in\mathcal{X}} p_x x.$$

This is just like the confusion in QM between operators and their eigenvalues.

Entropy as expected surprise. An incredibly useful functional of a probability distribution is the (Shannon) entropy

$$H[p] \equiv - \sum_{x\in\mathcal{X}} p_x \log p_x.$$

(We will normalize it with the log base two. And I will sometimes write square brackets to remind us that if we take a continuum limit of our sample space, then H is a functional.)



The quantity $-\log p_x$ can be called the *surprise* of x: if you know that the probability distribution is p_x, then you will be not at all surprised to get x if $p_x = 1$, and completely out of your mind if you got x when $p_x = 0$, and $-\log p_x$ smoothly interpolates between these values in between. So the entropy $H(X)$ is just

$$H[p] = \langle - \log p_x \rangle_X$$

the average surprise, or better, the *expected surprise.*

The entropy of a probability distribution measures how difficult it will be to predict the next outcome when sampling the distribution repeatedly. If we can make a simple rule for predicting the outcome, then we only need to keep track of the rule and its exceptions. This leads to the possibility of data compression (§2.2).

[Sethna §5.3.2] In case you think there is some arbitrariness in this choice of function, here are some (Shannon) axioms for a measure of ignorance:

1. Entropy is maximized for equal probabilities.

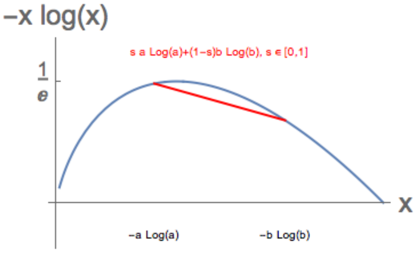
   This is true of $H[p]$ because $f(x) \equiv -x \log x$ is anti convex. This implies (let $\Omega \equiv |\mathcal{X}|$)

   $$\frac{1}{\Omega}\sum_k f(p_k) \leq f\left(\frac{1}{\Omega}\sum_k p_k\right) = f\left(\frac{1}{\Omega}\right).$$

   Multiplying the BHS by $-\Omega$ then says

   $$H[p] \leq H[u]$$

   where $u_x = \frac{1}{\Omega}$ is the uniform distribution.



2. Entropy is *stable* in the sense that adding extra states of zero probability doesn't change anything:

   $$H(p_1 ... p_\Omega) = H(p_1 ... p_\Omega, 0).$$

   This is true of $H[p]$ because $\lim_{x\to 0} x \log x = 0$.

3. Learning decreases ignorance (on average).

   More specifically, recall the notion of conditional probability. Suppose now that we have two discrete random variables A and B (with respective values A_n and B_l) with joint distribution $P(n,l) = \mathbf{Prob}(A_n \text{ and } B_l)$. The distribution for the second variable (ignoring the first) is

   $$q_l \equiv \sum_n P(n,l). \tag{2.1}$$

   (This is called a *marginal*.) The conditional probability for n given l is

   $$p(n|l) \equiv \frac{P(n,l)}{q_l}. \tag{2.2}$$

   (This is basically Bayes' rule. I'll say more about it below.) It is a normalized distribution for n, because of the definition of q_l (2.1).

   We can define a conditional entropy to quantify our knowledge of A given a value of B. If we measure B and find l, this is

   $$H(A|B_l) \equiv H(p(A|B_l))$$

where H is our entropy function. Its expected value, averaging over the result for B is then

$$H(A|B) \equiv \langle H(A|B_l) \rangle_B = \sum_l q_l H(A|B_l).$$

The third condition we want is: If we start with a joint distribution for AB and then measure B, our ignorance should decrease (on average) by our initial ignorance about B:

$$\langle H(A|B) \rangle_B = H(AB) - H(B).$$

Indeed this rule is satisfied by the Shannon entropy. That is:

$$\boxed{H(X,Y) = H(Y) + H(X|Y)}.$$

This boxed equation is called the chain rule. To prove it, just consider the log of Bayes' rule (2.2): $\log p(X,Y) = \log p(Y) + \log p(X|Y)$ and take $\langle \text{BHS} \rangle_{XY}$.

In particular, if A and B are uncorrelated, then $H(A|B_l) = H(A)$ for every l, and this rule says that we learn nothing and our ignorance doesn't change. More specifically, it says

$$H(AB) \overset{\text{uncorrelated}}{=} H(A) + H(B),$$

that the entropy is extensive in the case of uncorrelated subsystems.

The deviation from this condition is called the *mutual information*:

$$I(A:B) \equiv H(A) + H(B) - H(AB) = \sum_{ij} p(A_i, B_j) \log\left(\frac{p(A_i, B_j)}{p(A_i)p(B_j)} \right). \tag{2.3}$$

The argument of the log (which is sometimes called the *likelihood*) differs from 1 only if the two variables are correlated. It is a measure of how much we learn about A by measuring B.

The chain rule has various glorifications with many variables, *e.g.*:

$$H(X_1 \cdots X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1} \cdots X_1). \tag{2.4}$$

I am told that the previous three properties are uniquely satisfied by the Shannon entropy (up to the multiplicative normalization ambiguity). The basic uniqueness property is that the logarithm is the only function which satisfies $\log(xy) = \log(x) + \log(y)$. This comes in at desideratum 3.

Notice that the conditional entropy $H(A|B)$ is positive, since it's an average of entropies of distributions on A (each positive numbers). The chain rule then implies that $0 \leq H(A|B) = H(A,B) - H(A)$ so $H(A,B) \geq H(A)$. Since A isn't special, it's also bigger than $H(B)$ so it's bigger than the max of the two: $0 \leq \max(H(A), H(B)) \leq H(A,B)$.

Illustrations with inference problems. [Barnett §1.2; I highly recommend reading Chapter 3 of Mackay] Let's discuss some experiments with (for simplicity) two possible outcomes. I'll describe three different situations. In each case, our information about the situation is incomplete.

(1) In the first case, we know how often each outcome obtains. Let's say we're measuring some property of a physical system, call it property A which can be either $\uparrow$ or $\downarrow$, and we know that 1/4 of the time $A =\uparrow$: $p(A_\uparrow) = 1/4, p(A_\downarrow) = 3/4$. However, we have a very poor detector. It always says $\uparrow$ if $A =\uparrow$: $p(D_\uparrow|A_\uparrow) = 1$ but if $A =\downarrow$, it says $\downarrow$ only 3/4 of the time: $p(D_\downarrow|A_\downarrow) = 3/4$. The question is: if the detector says $\uparrow$, what probability should we assign to the statement that A is actually $\uparrow$?

The answer to this question is given by the thing that people usually call Bayes' rule, which is a rearrangement of (2.2) in the following form:

$$p(A_i|D_j) = \frac{p(D_j|A_i)p(A_i)}{p(D_j)}.$$

This is a distribution on outcomes for A, so we can use

$$p(A_i|D_j) \propto p(D_j|A_i)p(A_i)$$

and normalize later. In our example we have the numbers:

$$p(A_\uparrow|D_\uparrow) \propto p(D_\uparrow|A_\uparrow)p(A_\uparrow) = 1 \cdot \frac{1}{4}$$

$$p(A_\downarrow|D_\uparrow) \propto p(D_\uparrow|A_\downarrow)p(A_\downarrow) = \frac{1}{4} \cdot \frac{3}{4}$$

Since these have to add up to one and the second is 3/4 as big, we have $p(A_\uparrow|D_\uparrow) = 4/7$.

Suppose we measure twice the same configuration for A, independently, and get $\uparrow$ both times. Bayes rule generalizes to

$$p(A_i|D_j^1 D_k^2) = \frac{p(D_j^1 D_k^2|A_i)p(A_i)}{p(D_j^1 D_k^2)}$$

and we get a more certain outcome:

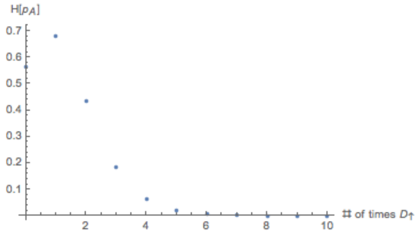
$$p(A_\uparrow|D_\uparrow^1 D_\uparrow^2) \propto \underbrace{p(D_\uparrow^1 D_\uparrow^2|A_\uparrow)}_{=p(D_\uparrow^1|A_\uparrow)p(D_\uparrow^1|A_\uparrow)} p(A_\uparrow) = 1 \cdot 1 \cdot \frac{1}{4}$$

$$p(A_\downarrow|D^1_\uparrow D^2_\uparrow) \propto p(D^1_\uparrow D^2_\uparrow|A_\downarrow)p(A_\downarrow) = \frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}$$

And we assign the detector being correct a probability of 16/19.

As we continue to measure ↑, the entropy in the distribution of our expectation for $A_\uparrow$ goes from

$$\begin{aligned} H(1/4,3/4) &= .56 \overset{D_\uparrow}{\to} \\ H(4/7,3/7) &= .68 \overset{D_\uparrow}{\to} \\ H(16/19,3/19) &= .44 \overset{D_\uparrow}{\to} \\ H(64/67,3/67) &= .18 \overset{D_\uparrow}{\to} \\ \ldots\, H\left(\frac{4^n}{3+4^n},\frac{3}{3+4^n}\right) &\overset{n\to\infty}{\to} 0. \end{aligned}$$



Exercise: How does $H(n) \equiv H\left(\frac{4^n}{3+4^n},\frac{3}{3+4^n}\right)$ decay as $n\to\infty$? This is a measure of how fast we learn.

(2) For the second example, suppose we are breeding *arctopuses*, diploid creatures used as a model organism by certain mad scientists, with two phenotypes: fire-breathing (↑) and not (↓). For better or worse, fire-breathing is recessive, so an arctopus with phenotype ↑ necessarily has genotype ↑↑, while a non-fire-breathing arctopus may be ↓↑, ↑↓ or ↓↓.

If we breed a firebreathing mother arctopus with a non-fire-breathing father, there are several possible outcomes. If the baby arctopus breathes fire then for sure the father was ↑↓ or ↓↑. If the offspring does not breathe fire then maybe the father was ↓↓. We would like to learn about the genotype of the father arctopus from observations of the progeny.

Unlike the previous problem, we don't know how often the three possibilities occur in the population (as you might imagine, arctopus genetics is a challenging field), so we must choose a *prior* distribution as an initial guess. Various forces argue for the maximum entropy distribution, where each possibility is equally likely:

$$p(\text{dad is }\downarrow\downarrow) = 1/3, \quad p(\text{dad is }\uparrow\downarrow\text{ or }\downarrow\uparrow) = 2/3.$$

(From now on I will not distinguish between ↑↓ and ↓↑ in the labelling.)

Now, if we repeatedly mate these arctopuses, we have

$$p(i\text{th offspring does not breathe fire}|\text{dad is }\downarrow\downarrow) = 1$$

$$p(i\text{th offspring does not breathe fire}|\text{dad is}\uparrow\downarrow) = 1/2.$$

If, as is likely, the first offspring does not breathe fire (I'll write this as $x_1 =\downarrow$), we infer

$$p(\text{dad is } \downarrow\downarrow | x_1 =\downarrow) \propto p(x_1 =\downarrow | \downarrow\downarrow) p(\downarrow\downarrow) = 1 \cdot \frac{1}{3}$$

$$p(\text{dad is } \uparrow\downarrow | x_1 =\downarrow) \propto p(x_1 =\downarrow | \uparrow\downarrow) p(\uparrow\downarrow) = \frac{1}{2} \cdot \frac{2}{3}$$

which when we normalize gives

$$p(\downarrow\downarrow | x_1 =\downarrow) = \frac{1}{2}, \quad p(\uparrow\downarrow | x_1 =\downarrow) = \frac{1}{2}.$$

If the second offspring also comes out ↓, we update again:

$$p(\downarrow\downarrow | x_1 =\downarrow, x_2 =\downarrow) \propto p(x_1 =\downarrow | \downarrow\downarrow) p(x_2 =\downarrow | \downarrow\downarrow) p(\downarrow\downarrow) = 1 \cdot 1 \cdot \frac{1}{3}$$

$$p(\uparrow\downarrow | x_1 =\downarrow, x_2 =\downarrow) \propto p(x_1 =\downarrow | \uparrow\downarrow) p(x_2 =\downarrow | \uparrow\downarrow) p(\uparrow\downarrow) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{3}$$

so now we assign $p(\downarrow\downarrow | ...) = 2/3$. We can think of this as updating our prior distribution based on new information.

Two comments:

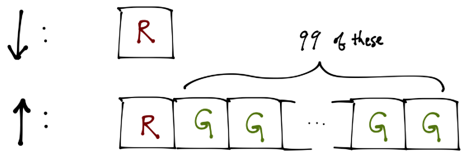
- The preceding examples should make clear that the probability we assign to an event are properties not just of the event, but also of our own state of knowledge. Given that I'm trying to persuade you in this class to think of a quantum state as a generalization of a probability distribution, you might worry that the same might be said about quantum states. This is an apocalypse-grade can of worms.

- Bayes' theorem is a theorem. It nevertheless carries with it a nimbus of controversy. The trouble comes from two parts: the first is the question of *interpretations of probability theory*, which is nearly isomorphic to its modern cousin *interpretations of quantum mechanics*. I don't want to talk about this.

  The second source of trouble is the assignment of prior distributions, and the choice of sample space for the prior. This is dangerous. Maximum entropy is great – it seems like it minimizes the introduction of unwarranted assumptions. However, the results it gives can depend on our assumptions about the space of possibilities. A sobering discussion for an ardent Bayesian is given in Aaronson's book, in the chapter called "Fun with anthropics", including the third example I can't resist discussing... [End of Lecture 3]

---

(3) The point of this example is to illustrate the point that one's theory of the world can affect the outcome of using Bayes' theorem. It is a puzzle due to Bostrom. Imagine a universe with a deity who flips a fair coin. If the coin says $\downarrow$, the deity makes one sealed room containing an intelligent person with red hair. If the coin says $\uparrow$ the deity makes 100 sealed rooms, each with an intelligent person. 99 of them have green-haired people and one has a red-haired person. Every room has a mirror and everyone knows the whole story I just told you.



If you wake up in a room and see you have green hair, then you know for sure the coin said $\uparrow$, $p(\downarrow|G) = 0$. The problem is: if your hair is red, what probability should you assign to $\uparrow$, *i.e.* what is $p(\uparrow|R)$? I emphasize that in the world of this story, this is a scientific question: the result of the coin flip is their version of the cosmic microwave background.

Theory A: Clearly it's a fair coin so the answer should be $\frac{1}{2}$, right? Bayes' rule says

$$p(\uparrow|R) = \frac{p(R|\uparrow)p(\uparrow)}{p(R)}$$

If the coin is $\downarrow$, then R is one possibility out of 100, so we conclude $p(R|\downarrow) = \frac{1}{100}$. A fair coin means $p(\uparrow) = \frac{1}{2}$. The denominator is

$$p(R) = p(R|\uparrow)p(\uparrow) + p(R|\downarrow)p(\downarrow) = 1 \cdot \frac{1}{2} + \frac{1}{100} \cdot \frac{1}{2} = \frac{1}{2} \cdot \frac{101}{100}.$$

So clearly

$$p(\uparrow|R) \overset{?}{=} \frac{1}{101}.$$

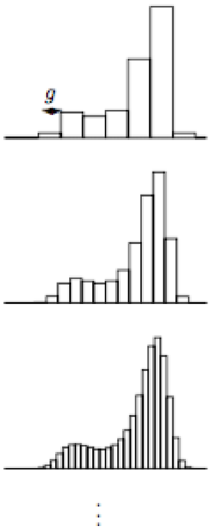
Theory B: There is another point of view. Suppose that the people take into account the information of their own existence. A person is much more likely to find themselves in a world with 100 people than a world with only 1 person, no? Only two people in a total of 101 people in the story have red hair, so clearly we must have $p(R) = \frac{2}{101}, p(G) = \frac{99}{101}$. In that case, you are more likely to find yourself in the $\downarrow$ world: $p(\downarrow) = \frac{100}{101}, p(\uparrow) = \frac{1}{101}$. Isn't it a fair coin? Yes, but here we are conditioning on the extra 'anthropic' information of finding ourselves to exist. In that case we get (it's still true that $p(R|\uparrow) = 1, p(R|\downarrow) = \frac{1}{100}$)

$$p(\uparrow|R) \overset{?}{=} \frac{\frac{1}{100} \cdot \frac{100}{101}}{\frac{2}{101}} = \frac{1}{2}.$$

So: while it's true that some properties of nature (the distance of the Earth from the Sun) are environmentally selected, probabilistic reasoning which conditions on our existence can be slippery.

More generally, the results of Bayesian reasoning depend on our theory of the world: on *which* sample space should we put the uniform prior? A related discussion in a more practical context is in this paper which I learned about from Roland Xu.

Comment on continuous distributions. I mentioned that I've been writing $H[p]$ in anticipation of the idea that the RV x could be continuous, so $p(x)$ would be a probability density, in which case the entropy becomes a functional $H[p] \overset{?}{=} \int dx\,(-p(x)\log p(x))$.



There are a few things that might bother you about this. First, a probability density is in general dimensionful (if x has dimensions), and it's bad karma to take the log of a dimensionful quantity. Even worse, we might want to arrive at a continuous distribution by approximating it by a family of discrete distributions with spacing Δx. But the Shannon entropies of those distributions actually approach

$$\int dx\,(-p(x)\log(p(x)\Delta x)) = \int dx\,(-p(x)\log p(x)) + \log\Delta x \overset{\Delta x\to 0}{\to} \infty.$$

For example, consider the case of the uniform distribution on an interval of length a. If we approximate this by $N = \frac{a}{\Delta x}$ points, we have $p_N(x_i) = \frac{1}{N}$ which has $H[p_N] = \log N$. [Fig is from Mackay, who calls $g \equiv \Delta x$.]

It's not surprising that there is a divergence in the entropy of a continuous distribution: the digits of a real number (with perfect precision) contain an infinite amount of information.

Fortunately, this headache is just an additive constant in the entropy. The mutual information (2.3), which is a difference of entropies, is perfectly well-behaved, and the factors of Δx (and the dimensions of the probability densities) all cancel in the argument of the logarithm. This lesson that the mutual information is 'UV finite' will be a good one to remember when we try to study entropies of states in field theories. Another UV-finite quantity is the ...

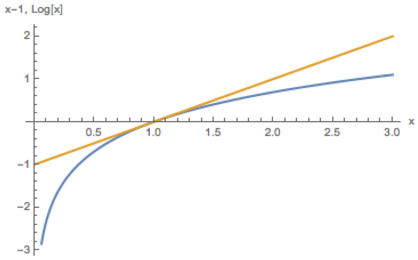
2.1 Relative entropy

Given two distributions p_x, q_x on the same random variable, their *relative entropy* is

$$D(p||q) \equiv \sum_{x\in\mathcal{X}} p_x \log\frac{p_x}{q_x}\ .$$

In the definition, samples $\alpha \in \mathcal{X}$ where $p_\alpha = 0$ don't contribute, but values where $q_\alpha = 0$ and $p_\alpha \neq 0$ give infinity. This quantity is sometimes called the 'Kullback-Leibler divergence'. Relative entropy is useful, and many of its properties generalize to QM. It is a sort of distance between distributions. It fails at this in some respects, for example because it is not symmetric in $p \leftrightarrow q$. [13]

Fact: $D(p||q) \geq 0$ for any p, q.



Proof: Let $A \subset \mathcal{X}$ be the support of p_x. The result follows from the fact that $\log x \leq x - 1$ for $x \in (0, \infty)$. (This is true because log is an anti-convex function on this domain (it is smooth and its second derivative is $-1/x^2 < 0$), so it lies *below* its tangents. The line $x - 1$ is tangent to $\log(x)$ at $x = 1$, as you can see in the figure; this is the only value which saturates the inequality.)

This means

$$\begin{aligned} -D(p||q) &= \sum_{x\in\mathcal{X}} p_x \log \frac{q_x}{p_x} = \sum_{x\in A} p_x \log \frac{q_x}{p_x} \\ &\leq \sum_{x\in A} p_x \left(\frac{q_x}{p_x} - 1\right) = \sum_{x\in A} (q_x - p_x) = \sum_{x\in A} q_x - 1 \leq 0. \end{aligned}$$

Equality only holds when $q = p$ (where $\log p/q = p/q - 1$). (Another proof of this statement uses Jensen's inequality: $-D(p||q) = \sum_{x\in A} p_x \log \frac{q_x}{p_x} \leq \log \sum_{x\in A} p_x \frac{q_x}{p_x}$.)

■

Relative entropy can be used to define the *mutual information* of two random variables $x \in X, y \in Y$ with joint distribution p_{xy} and marginals $p_x = \sum_{y\in Y} p_{xy}$ etc. (which we defined in (2.3)):

$$I(X:Y) \equiv D(p_{xy}||p_x p_y).$$

So the mutual info is a measure of distance to the uncorrelated case. (Beware the common abuse of notation I am making of denoting the distribution by the sample space, that is: the dependence on the choice of p_{xy} is implicit on the LHS.) Unpacking the definition,

$$\begin{aligned} I(X:Y) &= \sum_{xy} p_{xy} \log \frac{p_{xy}}{p_y p_x} = \left\langle \log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right\rangle_{XY} \\ &= -\sum_{xy} p_{xy} \log p_x + \sum_{xy} p_{xy} \log p(x|y) = H(X) - H(X|Y)\ . \end{aligned} \tag{2.5}$$

[13] So if we try to use the KL divergence to measure distance, p can be farther from q than q is from p. Emotional distance is a familiar example where such a thing is possible.

In red is Bayes' rule: $p(x|y) = \frac{p_{xy}}{p_y}$. This last expression allows us to interpret $I(X:Y)$ as the reduction in our uncertainty in X due to knowing Y. There was nothing special about singling out x in (2.5). It's also true that

$$I(X:Y) = -\sum_{xy} p_{xy} \log p_y + \sum_{xy} p_{xy} \log p(y|x) = H(Y) - H(Y|X) \ .$$

The case where $Y = X$ gives

$$I(X:X) = H(X) - \underbrace{H(X|X)}_{=0} = H(X)$$

which is why the entropy is sometimes intriguingly called the 'self-information'. Going back to the first expression, we can also recognize

$$I(X:Y) = H(X) + H(Y) - H(X,Y).$$

This follows from the chain rule $H(X,Y) = H(X) + H(Y|X)$.

An immediate consequence of our theorem that $D(p||q) \geq 0$ is

$$\boxed{I(X:Y) \geq 0}$$

since it is defined as the relative entropy of two distributions. And it vanishes only if the two variables are uncorrelated.

Another version of the same statement is *conditioning reduces entropy* (the third desideratum for H given above):

$$0 \leq I(X:Y) = H(X) - H(X|Y), \quad i.e. \quad \boxed{H(X) \geq H(X|Y)}.$$

Beware that this is a statement about the *average entropy* of X given Y. A particular value $H(X|Y=y)$ can be larger than $H(X)$, but $\sum_y p_y H(X|Y=y) \equiv H(X|Y) \leq H(X)$.

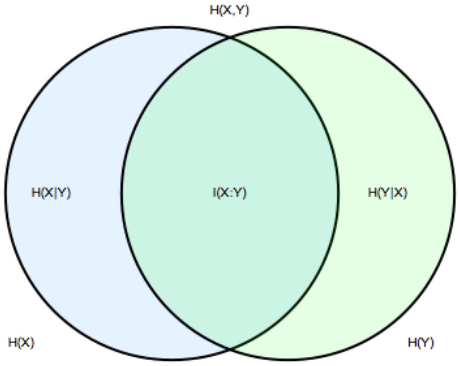
For example: consider the joint distribution $p_{yx} = \begin{pmatrix} 0 & a \\ b & b \end{pmatrix}_{yx}$, where $y =\uparrow, \downarrow$ is the row index and $x =\uparrow, \downarrow$ is the column index. Normalization implies $\sum_{xy} p_{xy} = a + 2b = 1$, so we have a one-parameter family of distributions, labelled by b. You can check that $H(X|Y) \leq H(X)$ and $H(Y|X) \leq H(Y)$ for any choice of b. However, I claim that as long as $b < \frac{1}{2}$, $H(X|Y=\downarrow) > H(X)$. (See the homework.)

The chain rule for H (2.4) then implies the "independence bound":

$$H(X_1 \cdots X_n) = \sum_{i=1}^{n} \underbrace{H(X_i|X_{i-1} \cdots X_1)}_{\leq H(X_i)} \leq \sum_{i=1}^{n} H(X_i)$$

which is saturated by the completely uncorrelated distribution $p_{x_1..x_n} = p_{x_1} \cdots p_{x_n}$. This is sometimes also called *subadditivity* of the entropy.

Here is a useful mnemonic[14]:



By the way, I said that two random variables (RVs) are uncorrelated if their mutual information vanishes. More generally, mutual information can be used to bound correlation functions, a representation of the amount of correlation between two RVs which is more familiar to physicists. In particular, given functions $\mathcal{O}_{X,Y}$ of random variables X, Y,

$$I(X:Y) \geq \frac{1}{2} \frac{\langle \mathcal{O}_X \mathcal{O}_Y \rangle_c^2}{\| \mathcal{O}_X \|^2 \| \mathcal{O}_Y \|^2}.$$

Here $\langle AB \rangle_c \equiv \langle AB \rangle - \langle A \rangle \langle B \rangle$ is the connected correlation function, and $\langle A \rangle \equiv \sum_{xy} p_{xy} A_x$. The norms in the denominator make it so multiplying our functions by some real number doesn't change the RHS; the definition is[15]

$$\| A \|^2 \equiv \sup_{p| \sum_x p_x = 1} \{ \sum_x p_x |A_x|^2 \}.$$

Later in §7.2, we'll prove the quantum version of this statement.

Next we will give some perspectives on why the Shannon entropy is an important and useful concept.

[14] There are some shortcomings of using a Venn diagram to illustrate entropies. I'll explain below in §2.3.1

[15] The definition of $\sup_{s \in S}\{f(s)\}$ here is the smallest number x such that $x \geq f(s), \forall s \in S$. Supremum differs from the maximum in that x need not be attained by $f(s)$ for any element of S.

2.2 Data compression

[Feynman, *Computation*, p. 121] The Shannon entropy of a distribution is sometimes called its 'information content' (for example by Feynman). In what sense does a random string of numbers have the largest information content? You learn the most about the next number (when you see it) if you have no way of anticipating it.

Why is $H(p) = -\sum_\alpha p_\alpha \log p_\alpha$ a good measure of the information gained by sampling the distribution p?

An example. [Mackay] Here is a demonstration that the surprise of an outcome $-\log p_{\text{outcome}}$ can usefully be regarded as the information gained by obtaining that outcome. In particular, you learn more when you obtain an improbable outcome.

Consider Mackay's more boring version of the game Battleship: A grid of $64 = 2^6$ squares contains one square occupied by a submarine. Each turn, the player guesses a square and is told whether it is a hit or miss.

Move number	outcome	p_{hit}	p_{miss}	info gained $= -\log p_{\text{outcome}}$	total info gained so far
1	miss	$\frac{1}{64}$	$\frac{63}{64}$	$-\log\frac{63}{64}$	$\log\frac{64}{63}$
2	miss	$\frac{1}{63}$	$\frac{62}{63}$	$-\log\frac{62}{63}$	$\log\frac{64}{63}\frac{63}{62} = \log\frac{64}{62}$
3	miss	$\frac{1}{62}$	$\frac{61}{62}$	$-\log\frac{61}{62}$	$\log\frac{64}{63}\frac{63}{62}\frac{62}{61} = \log\frac{64}{61}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
32	miss	$\frac{1}{33}$	$\frac{32}{33}$	$-\log\frac{32}{33}$	$\log\frac{64}{63}\frac{63}{62}\frac{62}{61}\cdots\frac{33}{32} = \log\frac{64}{32} = 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
48	miss	$\frac{1}{17}$	$\frac{16}{17}$	$-\log\frac{16}{17}$	$\log\frac{64}{16} = 2$
49	hit	$\frac{1}{16}$	$\frac{15}{16}$	$-\log\frac{1}{16}$	$2 + \log 16 = 6$

If you find the submarine on the first guess, the info gained is $-\log\frac{1}{64} = 6$ – you learned 6 bits of information from one yes/no question. This is because the outcome was very improbable. No matter when it happens, when you find the submarine, you have acquired 6 bits of information. In the game sequence in the table, why is the info gained from 32 consecutive misses equal to $\log 2 = 1$? Because by this point you've ruled out half the squares. That's equivalent to learning one binary digit of the location of the submarine (*e.g.* is the submarine in squares 1...32 or 33...64?).

Now for a more general viewpoint on why $H(p)$ is the average information gained by sampling p. Make a long list of samples from p, of length N: $x = \alpha_1\alpha_2\cdots\alpha_N$, which we'll think of as a message. (A useful notation: If α is a value of the RV A, then this x is an RV in A^N.) The number of appearances of a particular α is about Np_α. At large N we can ignore fluctuations about this average, and ignore the fact that Np_α need not be an integer. The number of *different* messages $\Omega(p)$ with this frequency distribution ($\equiv$ *typical messages*) is

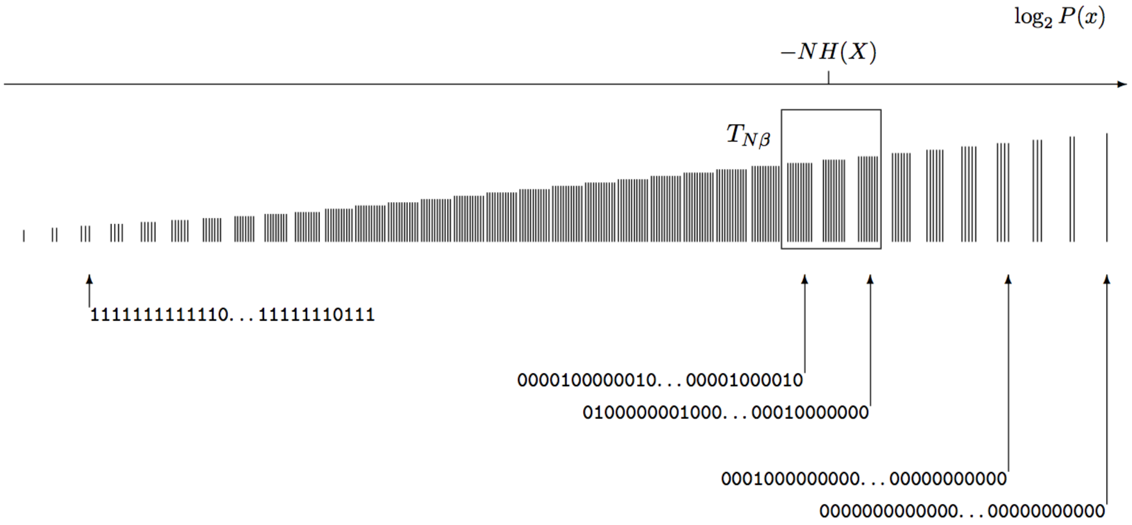
$$\Omega(p) = \frac{N!}{\prod_\alpha (Np_\alpha)!} \; .$$

Thinking of this as the number of microstates, the Boltzmann's-tomb, microcanonical notion of entropy is $\log \Omega$. Indeed, the "information expected per symbol" is

$$\begin{aligned} \frac{1}{N} \log \Omega &\overset{N \gg 1}{\simeq} \frac{1}{N} \left(N \log N - \sum_\alpha (Np_\alpha) \log (Np_\alpha) \right) \\ &= \; - \sum_\alpha p_\alpha \log p_\alpha = H(p). \end{aligned} \tag{2.6}$$

In the approximate step, we used Stirling's formula.

Notice that the single most probable message is in fact not in the typical set. To see this, here is a diagram from the great book by MacKay which I found illuminating:



He is studying a binary alphabet, with $p_0 > p_1 \equiv p$, and $P(x)$ is the probability of finding x, a particular string of N bits. The box contains the typical strings.

The crucial point is that the output is overwhelmingly likely to be a typical string. You should believe this if you believe the equipartition derivation of statistical mechanics (independently of whether you believe that derivation is relevant to why stat mech applies in the world). For the simple case of N iid random variables, the probability that a string x contains n ones is $p^n(1-p)^{N-n}$, which decays exponentially with n. The *number* of strings that contain n ones is $\binom{N}{n}$, which grows factorially in n. Therefore the number of 0s has a binomial distribution

$$P(n) = \binom{N}{n} p^n (1-p)^{N-n} \overset{N\gg 1}{\propto} e^{-\frac{(n-\langle n\rangle)^2}{2\sigma^2}} \qquad , \langle n\rangle = Np, \sigma = \sqrt{Np(1-p)}$$

which you know very well approaches a (narrow-width, $\frac{\langle n\rangle}{\sigma} \sim \frac{1}{\sqrt{N}}$) Gaussian at large N, by the central limit theorem.

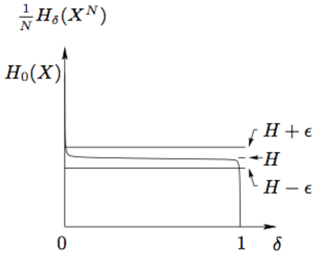
Since nearly all messages are typical, the number of bits we need to send in order to allow for the same number of different messages, is not N, but $NH(p)$.

The sketch I've just given can be made more precise by making an estimate of the errors from fluctuations about the average (rather than just ignoring them), and in that form is glorified (*e.g.* by Cover and Thomas) as the AEP (Asymptotic Equipartition Property). The more precise statement requires us to define the *essential bit content* of the RV X as follows: Rank the elements of the sample space $\mathcal{X}$ from most probable to least. Make a set S_δ by throwing in the elements of $\mathcal{X}$ starting from the most probable, until the total probability missing is δ. That is: S_δ be the smallest subset of the sample space $\mathcal{X}$ such that $P(x \in S_\delta) \geq 1-\delta$. This suggests a compression scheme where we assign codes to the elements of S_δ. The essential bit content is $H_\delta(X) = \log|S_\delta|$. A special case, where we allow no error, is $H_0 = \log|\mathcal{X}|$.

Then Shannon's noiseless-channel (or source) coding theorem says that given a RV X of entropy $H = H(X)$, and given $0 < \delta < 1, \epsilon > 0$, there exists a large-enough N so that

$$\frac{1}{N} H_\delta(X^N) < H + \epsilon \quad \text{and} \quad \frac{1}{N} H_\delta(X^N) > H - \epsilon.$$

The figure at right is from Mackay's book, which also provides a proof of this statement.



The conclusion is that we can use an alphabet with only 2^H symbols, which (if p is not uniform) is much smaller than $2^{H_0} = |\mathcal{X}|$ symbols. The first statement says that you don't need to use more than H symbols, and the second statement says that if you use any fewer you are guaranteed to miss some important information.

20 questions. [C&T p.110-112] Someone samples the distribution p_α and doesn't tell us which α results. We would like to formulate a series of yes/no ($\equiv 1/0$) questions which will uniquely and as-quickly-as-possible-on-average identify which α it is. The answers to the questions then comprise the binary digits of an efficient binary code for each element α in the sample set $\{\alpha\}$. Efficiency means minimizing the average code length

$$\langle \ell \rangle \equiv \sum_\alpha p_\alpha \ell_\alpha$$

where ℓ_α is the number of questions needed to identify uniquely element α.

Claim: The optimal $\langle \ell \rangle$ is $H[p]$. (This statement is equivalent to Shannon's source coding theorem since we can assign codewords to elements of the typical set.) If instead of binary, we used a D-symbol alphabet, we would have

$$\min \langle \ell \rangle = -\sum_\alpha p_\alpha \log_D p_\alpha \equiv H_D[p].$$

A strong interpretation of this statement, which is asymptotically correct, is: the optimal length of the codeword for symbol x should be its surprise.

The compression comes from using short sequences for common symbols: this is why the length should be the surprise. For example, consider the following table. For the distribution given in the table, $H = \frac{7}{4} = \langle \ell \rangle$. Notice that if such a compression scheme does not lose information (map multiple messages to the same code) then some (hopefully rare) messages must get longer.

Table 1

x	p_x	dumb code	Shannon optimal code	$-\log p_x$
A	$\frac{1}{2}$	00	0	1
B	$\frac{1}{4}$	01	10	2
C	$\frac{1}{8}$	10	110	2
D	$\frac{1}{8}$	11	111	3

Prefix codes and the Kraft inequality. A further demand we might make, for example, if we were interested in using this code to send messages using the alphabet $\{\alpha\}$, is that the code be a *prefix code*, which means that you can tell when a codeword ends – no two code words begin the same way. (Synonyms are *instantaneous* or *self-punctuating*, since you can tell right away when a new codeword starts.) Such a code works like a binary tree, beginning at the left from the first question and going up or down depending on the answer to each question. Efficiency means that some branches of the tree end early, before $\ell_{\max}$ questions, thereby removing all the potential daughter leaves. A codeword of length ℓ eliminates $D^{\ell_{\max}-\ell}$ terminating daughter leaves (at depth $\ell_{\max}$). The number of terminating leaves of the tree which are not codewords is then

$$\sum_\alpha D^{\ell_{\max}-\ell_\alpha} \leq D^{\ell_{\max}}$$

where $D = 2$ for a binary tree. Dividing the BHS by $D^{\ell_{\max}}$ then gives the Kraft inequality

$$\sum_\alpha D^{-\ell_\alpha} \le 1. \tag{2.7}$$

You might think that a prefix code is a strong demand. A code which you can concatenate without ambiguity (but you maybe can't tell until the end how to parse) it is called *uniquely decodable*. (That is: a code X is uniquely decodable if X^N is *not singular*, where singular means two plaintext messages map to the same codeword.) Kraft's theorem actually says a stronger thing, namely that for any uniquely decodable code there exists a prefix code with the same $\langle \ell \rangle$ (and we already showed that this inquality holds for prefix codes). [End of Lecture 4]

Here's why [C&T p.116-117]: Consider

$$\left(\sum_{x\in X} D^{-\ell_x}\right)^k = \sum_{x_1\cdots x_k\in\mathcal{X}^k} D^{-\sum_{i=1}^k \ell(x_i)}$$

and gather the terms by total length, m:

$$= \sum_{m=1}^{k\ell_{\max}} \underbrace{a(m)}_{\le D^m} D^{-m} \le k\ell_{\max}.$$

The number of sequences in a segment of length m in a D-ary code is D^m, and unique decodability means they can't appear more than once. So $\forall k$,

$$\sum_{x\in X} D^{-\ell_x} \le (k\ell_{\max})^{1/k} \overset{k\to\infty}{\to} 1.$$

So there are just as many prefix codes as uniquely decodable codes: no need to wait until the end of the message to start parsing.

Here's a physics proof that $H(p)$ is the optimal number of questions, *i.e.* the optimal average length of a prefix code. Minimize $\langle \ell \rangle = \sum_\alpha p_\alpha \ell_\alpha$ subject to the Kraft inequality (2.7).

We can do pretty well by ignoring the constraint that ℓ_α are integers and assuming (2.7) is saturated, imposing it with a Lagrange multiplier λ:

$$J[\ell_\alpha] \equiv \sum_\alpha p_\alpha \ell_\alpha + \lambda \left(\sum_\alpha D^{-\ell_\alpha} - 1\right)$$

is extremized when

$$0 = \partial_{\ell_\alpha} J|_{\ell=\ell^\star} = p_\alpha - \lambda \log D D^{-\ell^\star_\alpha} \implies D^{-\ell^\star_\alpha} = \frac{p_\alpha}{\lambda \log D}$$

but the constraint determines $1 = \sum_\alpha D^{-\ell_\alpha} = \frac{1}{\lambda \log D} \sum p_\alpha = \frac{1}{\lambda \log D}$ so we get $\ell^\star_\alpha = -\log_D p_\alpha$ and

$$\langle \ell \rangle_\star = \sum_\alpha p_\alpha \ell^\star_\alpha = -\sum_\alpha p_\alpha \log_D p_\alpha = H_D(p).$$

And the extremum is actually a minimum: $\boxed{\langle \ell \rangle \geq H_D[p]}$. To see this, notice that $q_\alpha \equiv \frac{D^{-\ell_\alpha}}{\sum_\beta D^{-\ell_\beta}}$ is a possible distribution on code lengths. Now consider the difference

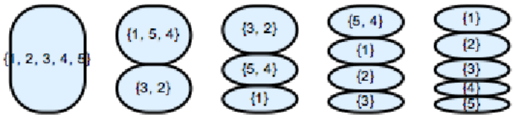
$$\begin{aligned}\langle \ell \rangle - H_D(p) &= \sum_\alpha p_\alpha \underbrace{\ell_\alpha}_{=\log_D\left(D^{\ell_\alpha}\right)} + \sum_\alpha p_\alpha \log_D p_\alpha \\ &= \underbrace{\sum_\alpha p_\alpha \log_D\left(\frac{p_\alpha}{q_\alpha}\right)}_{\equiv D(p||q) \geq 0} + \underbrace{-\log_D\left(\underbrace{\sum_\alpha D^{-\ell_\alpha}}_{\overset{(2.7)}{\leq} 1}\right)}_{\geq 0} \end{aligned} \tag{2.8}$$

Here $D(p||q)$ is the relative entropy.

Huffman codes and strong-disorder RG.

The preceding discussion does nothing to help us find a good code. An optimal binary symbol code can be made by the following 'greedy' recursive procedure: Order the elements by their probability. First group the two least probable outcomes p_n, p_{n-1} into one element of a smaller sample set. Their codewords will only differ in the last digit. The smaller sample set has one fewer element – instead of p_n, p_{n-1} we have just the composite element with probability $\tilde{p}_{n-1} = p_n + p_{n-1}$. Repeat. Codewords only acquire a digit at the coarse-graining step (I'm using the convention that the less probable element gets a 1). An example will help a lot: [16]

x	p_x					codeword
1	.25	.3	.45	.55	1	01
2	.25	.25	.3	.45		10
3	.2	.25	.25			11
4	.15	.2				000
5	.15					001



$\longleftarrow$ This way to the IR.

[16] Warning: there is another way to implement the procedure, used by Mackay, which will result

In this code, the average string length is 2.3; the entropy of the distribution is 2.28548.

(For a brief introduction to strong-disorder RG, see the discussion in the last section of my 217 notes.)

Huffman coding produces an optimal *symbol code*, where we make a code for each letter of an alphabet. This takes no advantage of correlations between successive letters in the stream of symbols. A code which does is called a *stream code*, for which see chapter 6 of Mackay. Such codes are used by gzip and lots of other common compression applications. One example is called arithmetic coding, the basic idea of which comes from the following.

Shannon code. [C&T Problem 5.5.28] Where did those numbers come from in the Shannon optimal code I showed in Table 1? Put the objects in order of probability. Consider the cumulative probabilities $F_i = \sum_{j=0}^{i-1} p_i$. We can associate each element with the range of numbers $[F_i, F_{i+1})$. To uniquely specify the *ith* interval, we need to keep $\lceil \log 1/p_i \rceil$ digits of F_i. So those codewords in the table are actually just the first digits of F_i. By construction, the average codeword length is $\langle l \rangle = \sum_i p_i \lceil \log 1/p_i \rceil \leq H(X) + 1$. Because the p_i happen to be powers of two, in this case $\lceil \log 1/p_i \rceil = -\log p_i$ and $\langle \ell \rangle = H(X)$. Notice that $\ell_i = \lceil \log_D 1/p_i \rceil$ is long enough for a prefix code, since $\sum_i D^{-\lceil \log_D 1/p_i \rceil} \leq \sum_i D^{\log_D p_i} = \sum_i p_i = 1$ satisfies the Kraft inequality. And this does give a prefix code because the intervals don't overlap – once we have enough digits to specify the interval, we can't possibly have those same digits for any other interval.

The wrong code. What if we think the distribution is q_x but in fact it's p_x, and we make an optimal code for q_x? The expected length is

$$\langle \ell_q \rangle_p \simeq \sum_x p_x (\log \frac{1}{q_x}) = \sum_x p_x \log \frac{p_x}{q_x p_x} = D(p||q) + H(p).$$

(More precisely, the LHS can be bounded between this number and this number plus one.) This gives another interpretation of the relative entropy.

2.3 Noisy channels

[Barnett §1.4] We can put the previous discussion into the context of the theory of communication: the goal is to transmit information (through space or time). This

in a different code (but the same word-lengths). The difference is that Mackay doesn't sort the list after the first step. I prefer to sort the symbols, so that the definition is recursive i.e. its the same algorithm at each step.

process is necessarily probabilistic, since if the receiver knew for sure what the message was, there would be no point.

The sender is a random variable called A and the receiver is a random variable called B. A *channel* is characterized by $\{p(b|a)\}$ a set of probabilities for the receiver to get b when the sender sent a. B would like to know $p(a|b)$. We suppose a distribution $p(a)$ on A, known to B for example by previous interaction through the channel.

If $p(a|b) = \delta_{ab}$, then the channel is as good as can be, and this was what we supposed in the last subsection. Now we introduce *noise*.

Notice that communication across space is not the only relevant context for this discussion: *memory* is communication through time. For example, by writing something down on a piece of paper, I can communicate with myself in the future. On the other hand, I may lose the piece of paper ...

2.3.1 Binary symmetric channel

[MacKay, exercise 8.7 and 8.8] Consider three correlated random variables, A, E, B. Think of A as the sender, B as the receiver and E as a source of noise. They are all binary variables. We'll take A and E to be independent, with $p(a) \equiv (1-p, p)_a$, $p(e) \equiv (1-q, q)_e$. A and E jointly determine the result of B to be

$$b = (a+e)_2 \equiv (a+e) \text{ modulo } 2.$$

So $e = 0, 1$ code for 'no error' and 'yes error', and a bit-flip error happens with probability q.

Notice that if $q = \frac{1}{2}$ – a bit flip is as likely as not, then A and B are completely uncorrelated: $I(A:B) = 0$.

However: if we know the value of the noise bit (whatever it is), A and B are perfectly correlated. This is a good opportunity to introduce the *conditional mutual information*. Just like the mutual information, it is best defined using the relative entropy:

$$I(A:B|E) \equiv D(p(AB|E)||p(A|E)p(B|E))$$

which shows that it is positive. It is also just $I(A:B|E) = H(A|E) - H(A|BE)$. [17]

[17] Notice that I sometimes drop the commas between the random variables; notice also that the comma is more powerful than the | or the :, so that for example $H(A|BE)$ means $H(A|(BE))$.

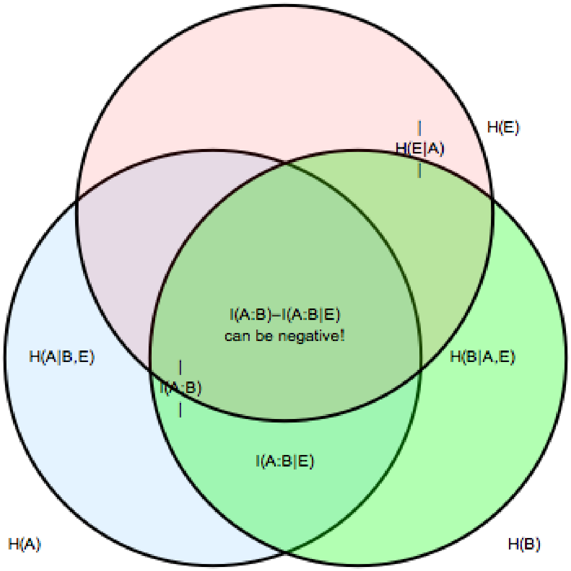
But now consider the example above, for simplicity in the case with $q = \frac{1}{2}$, so that $I(A:B) = 0$. But the conditional mutual information quantifies our statement that if we measure the noise then we restore the correlation between A and B:

$$I(A:B|E) = H_2(p) > 0.$$

This means that the area in the central region of the figure below is actually negative. The diagram is not wrong, but we must not interpret it too literally.
It does correctly predict relations like

$$H(ABE) = H(A) + H(E|A) + H(B|A,E) \tag{2.9}$$

which follows from the chain rule.

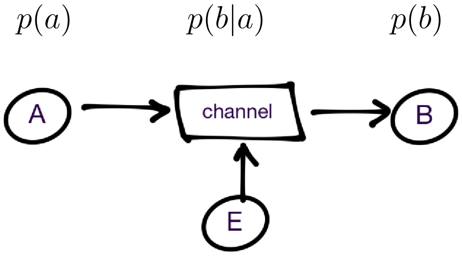


2.3.2 Noisy channel Shannon theorem

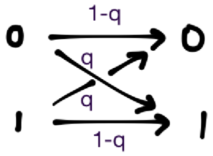
In the previous subsection, redundancy in our messages was a nuisance which we wanted to remove to more efficiently use our wonderful clean channel. Here we consider the case where the channel is noisy and we wish to ask how much redundancy is needed to protect the message against noise.

To see that redundancy can protect against noise, notice tht t s stll pssbl t rd ths sntnc vn thgh ll th vwls hv bn rmvd. English is very highly redundant. In fact, even though it nominally uses a 26-letter alphabet (potentially almost 8 bits), it is estimated to convey (by an experiment designed and performed by Shannon!) only about one bit per letter. Part of this is the non-uniform distribution of the letter frequencies (see HW 3), and also of the frequencies of 2-, 3- and more letter combinations. But part of it is semantic: neighboring words are quite strongly correlated. So, in general, you can often predict pretty well what the next letter will be if you watch someone typing in English. (See C&T §6.4 for a great discussion of the entropy of English.) This ability to predict the future well means that you can also compress the signal well. (It is also equivalent to being able to take advantage of gambling opportunities.) This perspective leads to compression algorithms better than any symbol code (of which the Huffman code is optimal).

Now let's go back to our noisy channel, and suppose we've already optimally compressed our message of 2^{N_0} bits. So we choose from 2^{N_0} messages of equal probability. In the picture of the channel at right, we assume that B has no direct knowledge of E. (Note that E is for 'environment'.) So the channel is characterized by $p(B|A)$ – it determines probabilities for what comes out, according to what went in.



The binary symmetric channel described above simply says that each bit sent can be flipped with probability q. (We drop the assumption that successive source bits A are uncorrelated.) On average, then, qN_0 wrong bits will be received. Again, the distribution of the amount of wrongness is very sharply peaked at large N_0.



To fix the errors, B needs to know *which* bits are wrong. For a typical message, there are

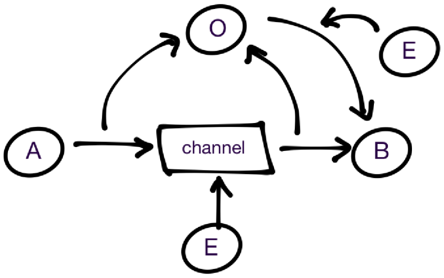
$$N_E = \frac{N_0!}{(qN_0)!((1-q)N_0)!}$$

ways of distributing the qN_0 errors among the message bits. So, to specify their locations, B needs

$$\log N_E \overset{\text{Stirling}}{\simeq} N_0 H(q)$$

extra bits of information.

Suppose an all-seeing observer looks at the received bits and compares them with the correct ones; such an observer would need to send B an extra $N_0H(q)$ bits, so B gets $N_0(1 + H(q))$ bits.
But suppose further that the all-seeing observer must also use the same noisy channel (a burning bush, say) with error rate q per bit.



We need to correct the errors in the $N_0H(q)$ correction bits; that takes an extra $(N_0H(q))H(q) = N_0H(q)^2$ bits. And of course we can't stop there; altogether B must receive

$$N = \sum_{k=0}^{\infty} N_0 H(q)^k = \frac{N_0}{1 - H(q)}$$

total bits to get the message through the noisy channel.

Why did we use the same q for the omniscient-observer phone? Because then we can just use this result to describe what happens when A herself sends the corrections!

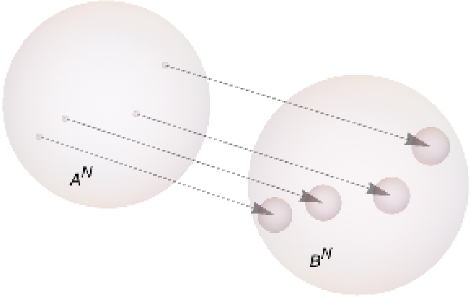
So the right way to think about this is that N bits sent through a noisy channel encode only

$$2^{N_0} = 2^{N(1-H(q))} \text{ distinct messages.}$$

Each transmitted bit carries only

$$\frac{1}{N}\log\left(2^{N(1-H(q))}\right) = 1 - H(q) \text{ bits of information.}$$

Where does this reduction in efficacy (I guess the right word is 'capacity') of a noisy channel come from? Each message sent gets scrambled away from its target to a typical set of $2^{NH(q)}$ received messages. Think of this as a ball (of a radius determined by the error rate) around the intended message in the space of messages. In order for these messages to be distinguishable from each other, A has to send only *sufficiently different* messages. Sufficiently different means their error balls don't touch, so there are only $2^{N(1-H(q))}$ such messages we can pack in there.



[End of Lecture 5]

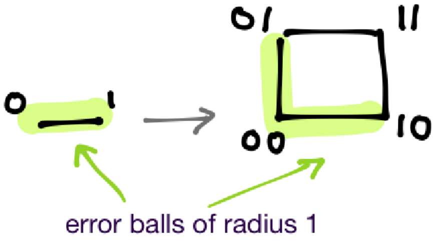
Hamming distance. What is the distance measure we are using on the space of messages (which is pink) in the lovely figure above? A convenient one, which changes by 1 each time a bit is flipped is the *Hamming distance* which for two binary strings of length N is

$$d_H(x,y) \equiv \sum_{\text{digits},i=1}^{N} (x_i - y_i)_2 = \text{the \# of digits which differ.}$$

Related concepts are Manhattan distance and trace distance. (Beware that for non-binary (*e.g.* ternary variables) people still define the Hamming distance to be the number of digits which differ.) This quantity *is* a distance: it is positive, and only vanishes if $x = y$, it is symmetric under interchange of x, y, and it satisfies the triangle inequality $d_H(x,y) \leq d_H(x,z) + d_H(z,y)$.

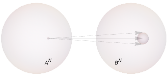
So e (distinct) errors move the target message a distance e. It is a random walk on a hypercube of e steps, starting at the correct message. The minimum distance d_H ($\equiv d$) between codewords determines B's ability to detect and correct errors. In particular B can detect $d-1$ errors and correct $\frac{1}{2}(d-1)$. Whence these numbers: Until there are d errors, a message can't make it all the way to another codeword. And until there are more than $\frac{1}{2}(d-1)$ errors, the message is closest to the correct codeword than any other.

In this language, a repetition code works because of Pythagoras (or rather the Pythagoras of Manhattan): The distance between 0 and 1 is 1, but the distance between 00 and 11 is 2.

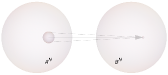


There are better ways to do this, better in the sense that the length of the message need not grow so quickly with the amount of error-protection that results. More on this below in §2.4.

Channel capacity. So A has $2^{NH(A)}$ typical messages to choose from to send. Each sent message produces $2^{NH(B|A)}$ received messages.



B has $2^{NH(B)}$ typical messages to choose from to receive. Each received message is produced by $2^{NH(A|B)}$ sent messages.



(These are like forward and backward light cones in the message space.) So the

$$\#\text{ of reliably sendable messages } = 2^{N(H(B)-H(B|A))} = 2^{N(H(A)-H(A|B))} = 2^{NI(A:B)} \ .$$

The equals signs here are in the sense of the AEP and become exact at large N. The mutual information determines how much information can be sent. Yay, the mutual information.

This is not yet a property of the channel, since A has some discretion about her distribution. The *channel capacity* extremizes over this freedom

$$C \equiv \sup_{p(A)} I(A:B) \ .$$

In the supremum here, we vary $p(a)$, fixing $p(b|a)$. 2^{NC} is then the best number of messages A can send with N symbols by changing her strategy for weighting them.

For example, for the binary symmetric channel,

$$p(b|a) = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix}_{ab}$$

and $p(ab) = p(b|a)p(a)$ where $p(a)$ is to be determined. Now for simplicity we'll put back our assumption of uncorrelated successive bits from A, and let $p(0) = p$. So

$$I(A:B) = -\sum_{ab} p(ab) \log\left(\frac{p(ab)}{p(a)p(b)}\right) = H(A) - H(A|B) = H_2(q) - H_2((p(1-q)+(1-p)q))$$

is maximized when $p = \frac{1}{2}$, and the capacity is $C = 1 - H(q)$.

2.4 Error-correcting codes

It is not our business to do too good a job at this, but some of the ideas and language will be useful later. In particular, there is a close connection to the physics of topological order.

Suppose we want to send a string of bits $a_1 \cdots a_N$ through a noisy channel. If we send instead one extra bit (say, at the beginning), $a_0 a_1 \cdots a_N$, where $a_0 = (\sum_{i=1}^N a_i)_2$ (and the receiver knows we're doing this), then (at the cost of just one extra bit) the receiver can detect (but not locate) whether there has been an (odd number of) error(s). He just has to check the parity of the sum of the message bits against a_0.

If instead we arrange our bits into an $n \times m$ grid a_i^j,

$$\begin{pmatrix} a_1^1 & \cdots & a_1^m & \left(\sum_j a_1^j\right)_2 \\ a_2^1 & \cdots & a_2^m & \left(\sum_j a_2^j\right)_2 \\ \cdots & \ddots & \vdots & \vdots \\ a_n^1 & \cdots & a_n^m & \left(\sum_j a_n^j\right)_2 \\ (\sum_i a_i^1)_2 & \cdots & (\sum_i a_i^n)_2 & \left(\sum_{ij} a_i^j\right)_2 \end{pmatrix}$$

we can locate a single error by identifying which rows and columns disagree with their parity-check bits. The lower right corner allows us to check our checks, so we can identify whether there are two errors.

This is an example of a **Hamming code**. The bits to transmit are determined by a linear function of the message bits.

Here's a more systematic example: a '[7,4] Hamming code' encodes uses 7 transmitted bits to send 4 logical (message) bits as follows: To encode the message

$$s = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix}, \qquad \text{send} \qquad t = \begin{pmatrix} \mathbb{1}_{4\times 4} \\ P \end{pmatrix} s \equiv \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} s \equiv Gs$$

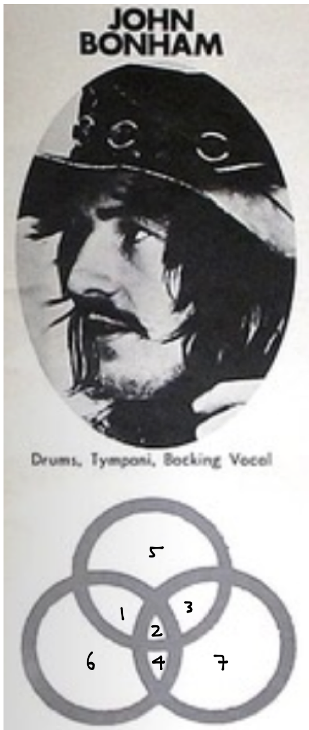
(the equality should be understood mod 2, and missing entries are zero).

An alternative equivalent way to define the codewords is by the condition that $t_1 + t_2 + t_3 + t_5$ is even, $t_1 + t_2 + t_4 + t_6$ is even, and $t_2 + t_3 + t_4 + t_7$ is even, *i.e.* $Ht = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ mod 2, where

$$H \equiv (P|\mathbb{1}_{3\times 3}) \,.$$

(*i.e.* $HG = 0$ mod 2). These conditions are like the parity-check conditions.

The decoder then acts on the received message $r = t + n$ (a 7-component column, where n is the noise) by the (partial) inverse map H. Since $Ht = 0$ for codewords, anything that gets through is noise: the *syndrome* is $z = Hr = Hn$. Since each s appears in two parity checks, the syndrome can detect two errors (and correct one). The receiver then reconstructs the message by finding the smallest number of errors which account for the syndrome. A useful mnemonic for the [7,4] Hamming code, popularized by Led Zeppelin, appears at right. The circles represent the three parity checks; each message bit, 1-4, is inside two of the circles.



Brief preview of the connection to physics. Consider a classical spin system made of 7 spins $Z_i = (-1)^{t_i}$ ($t_i = 0, 1$ means $Z_i = \pm 1$). Consider the Hamiltonian

$$H = -Z_1 Z_2 Z_3 Z_5 - Z_1 Z_2 Z_4 Z_6 - Z_2 Z_3 Z_4 Z_7 .$$

The low-energy subspace of this model is exactly the codewords of the [7,4] Hamming code.

Rat code. How does the number of parity check bits scale with the number of message bits? On HW3, there is a problem with 7 rats which are used to locate poison in (at most) one of 127 vials of liquid. Vials of liquid are like message bits, $s_i, i = 1..127$ and rats are parity check bits, $n = 1..7$. Here's the code:

$$G = \begin{pmatrix} \mathbb{1}_{127\times 127} \\ f_{i,n} \end{pmatrix}, \quad f_{i,n} = \begin{cases} 1 & \text{if rat } n \text{ drinks from vial } i \text{ (in your solution to the rat problem)} \\ 0 & \text{if not} \end{cases} .$$

For the same reason that your solution to the rat problem locates the poison, this code will locate a single error. This is an argument that to locate a single error, the number of parity check bits should scale like the log of the number of message bits.

The design of good error correcting codes is a huge industry. The rat code is the beginning of the story of a family called Reed-Muller codes. One measure of good is many logical bits k encoded in few raw bits N (which were $N = 7, k = 4$ for the example above). Another desideratum is a large code distance ($\equiv$ minimum distance between code words). The subject has deep connections to sphere packing (perhaps not surprising given the picture described above) and to sporadic finite groups[18] An interesting question is: can we saturate the Shannon bound (for a channel with an amount of noise so the average number of errors per block is below the code distance)? The answer was 'no' for a long time, and the codes which do so are a bit more involved. The ones used in your CD player (if you still have one) are like stream codes, in that the codewords depend on past input. A comprehensive book on this subject is the one by W. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes.*

End-of-act-one discouragement by way of preview. Consider for a moment the quantum version of the above ideas: we have some precious quantum state which we want to send down a noisy channel to our friend Bob. There are many reasons to be discouraged about the prospects for doing this:

(1) Say our message state is a single-qbit pure state $|\psi\rangle = z|0\rangle + w|1\rangle$, $z, w \in \mathbb{C}$. We could try to send the two real numbers which specify the point on Bloch sphere. A priori, this isn't such a great idea, since a single real number has infinitely many bits. And you can see that this probably isn't on the right track since when we want to send larger states, say of N qbits, we would need to confront the Illusion of Hilbert Space, with its 2^N complex numbers, head-on.

(2) Quantumly, here are many more possible ways things can go wrong. For example, in addition to bit-flip errors, where a $|0\rangle$ is replaced by a $|1\rangle$, we can also get the phase wrong, *e.g.* a transmitted $|\psi\rangle$ could become $z|0\rangle - w|1\rangle$. Or even some (gasp) continuous variation of the phase.

(3) So we'll need to learn to correct these errors. But notice that both repetition codes and parity-check codes involve ingredients which are hard (meaning: either fraught or simply impossible) to do in quantum mechanics, namely *copying* and *measurement*. Furthermore, I've been speaking as if we *know* the complex numbers z, w. But we certainly cannot determine those from a single copy of the state $|\psi\rangle$.

No cloning fact. Why can't we copy a quantum state? Suppose we have a unitary map which for any (unknown) state $|a\rangle$ acts by

$$\textbf{Xerox} : |a\rangle \otimes |\text{anything}\rangle \mapsto |a\rangle \otimes |a\rangle \ .$$

[18] For more on these connections I recommend the book by Thomas Thompson, *From error correcting codes through sphere packings to simple groups.*

If it's supposed to copy any state, then similarly we must have

$$\mathbf{Xerox}\,|b\rangle \otimes |\text{anything}\rangle = |b\rangle \otimes |b\rangle \ .$$

But then what does it do to the superposition? By assumption, it copies it:

$$\mathbf{Xerox}\left(\frac{|a\rangle + |b\rangle}{\sqrt{2}} \otimes |\text{anything}\rangle\right) = \left(\frac{|a\rangle + |b\rangle}{\sqrt{2}}\right) \otimes \left(\frac{|a\rangle + |b\rangle}{\sqrt{2}}\right) \ .$$

But that's not the same as the superposition of the images:

$$\begin{aligned}\mathbf{Xerox}\left(\frac{|a\rangle + |b\rangle}{\sqrt{2}} \otimes |x\rangle\right) &\neq \frac{1}{\sqrt{2}}\left(|a\rangle \otimes |a\rangle + |b\rangle \otimes |b\rangle\right) \\ &= \frac{1}{\sqrt{2}}\left(\mathbf{Xerox}\,|a\rangle \otimes |x\rangle + \mathbf{Xerox}\,|b\rangle \otimes |x\rangle\right).\end{aligned}$$

So such a map as **Xerox** can't even be linear, never mind unitary. (Why can't we make a machine that does nonlinear operations on quantum states? Machines that I know about act by time evolution using some Hamiltonian governing the dynamics of the constituents. You might imagine that open quantum systems evolve by some more mysterious evolution, but in fact their time evolution too can be derived (by the Stinespring dilation theorem, about which more later) from unitary evolution on a larger Hilbert space. If you find a way to violate linearity of quantum mechanics, tell me and no one else. Here are some examples of things that go wrong.)

So you can find operators that copy specific known states, but never arbitrary superpositions. Note that there is a clever workaround for *moving* quantum information, which is cynically called *quantum teleportation*. This is a protocol to *move* an unknown quantum state of a qbit (from one tensor factor of $\mathcal{H}$ to another), by sending two classical bits, using some entanglement as lubricant. However, only one copy of the unknown quantum state is present at any time.

So the no-cloning fact is a serious obstacle to making 'quantum repetition codes'. Similarly, it sure seems like a 'quantum parity check code' would require us to measure the state (in some basis) so that we can determine the parity check bits. But measuring some observable acting on a quantum state is notorious for disturbing that state.

Amazingly, all of these problems have been overcome in the theory of quantum error correction. And you can understand many of the results in this area if you understand the toric code Hamiltonian. This will be the subject of §8.

3 Information is physical

The basic point is this. The following two situations are quite distinct from the perspective of thermodynamics: In situation A, we have a box of gas with average energy NT. In situation B, we have a box of gas with average energy NT and we know that all the molecules are on the left side of the box. Notice that I say 'situations' and not 'states' because the way in which A in B differ is a property of our knowledge, not of the atoms in the box.

These two situations have very different free energy F and entropy S, $F = E - TS$. Why should we care about that? In case B we can take advantage of our knowledge to do work: we can place a partition to keep the atoms on the left side, and then we can let the gas expand against the partition (say reversibly, at constant temperature), extracting heat from the bath and doing useful work on the partition.

Quantitatively, let's assume an ideal gas (so that E is independent of V) and

$$\Delta F|_{\text{fixed }T} = \underbrace{\Delta E}_{=0} - T\Delta S = -T\Delta S \ .$$

The heat *extracted* from the bath during the expansion, ΔQ satisfies $\Delta Q \geq T\Delta S$, and the inequality is saturated if the expansion is done reversibly. The entropy change of the system is $\Delta S = Nk_B \ln\left(V_2/V_1\right) = Nk_B \ln 2$.

Exactly because of this entropy difference, situation B sounds very unlikely for a large number of molecules, so who cares about this? In response to that, let us boldly set $N = 1$. Then the entropy difference is just one bit (or in thermodynamics units, it is $k_B \ln 2$).

You might be bothered by the idea of a one-molecule ideal gas. You should not be too bothered. Here are two reasons it is OK: One reason it is OK is that we can time average. The second, better reason is that the equilibrium thermodynamics of a single free particle is perfectly well-defined, even classically:

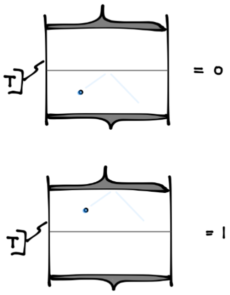
$$Z_1 = \int d^d p d^d q e^{-\beta \frac{p^2}{2m}} \propto T^{d/2} V, \ \ F = -k_B T \ln Z = -k_B T \left(\ln V + \frac{d}{2} \ln T \right) .$$

The walls of the container can keep the particle in equilibrium.

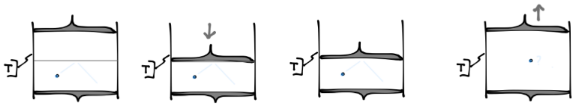
3.1 Cost of erasure

[Plenio-Vitelli, quant-ph/0103108; Barnett, §1.4; Feynman *Lectures on Physics*, Volume I, §46; Feynman *Lectures on Computation*, chapter 5; Sethna, chapter 5, especially problem 5.2; Bennett, *The thermodynamics of computation – a review*]

Pushing this idea a bit further, we can make a one-bit memory out of our one-atom ideal gas. The doohickey on the left of the figure is a contact with a heat reservoir at temperature T. There is a removable partition separating the two sides, and the top and bottom are frictionless pistons which may be attached to a weight machine to do work.



Burning information as fuel. Consider the diagrams at right. If you know the value of the bit (for example, look in the box), you can use it to do work, as in the diagrams. (If the value of the bit is 1 instead of 0, the process must be adjusted accordingly.) This is the same process as in the opening paragraph of this section.

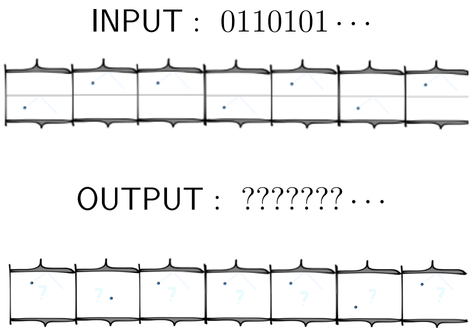


The gas does work *on* the piston

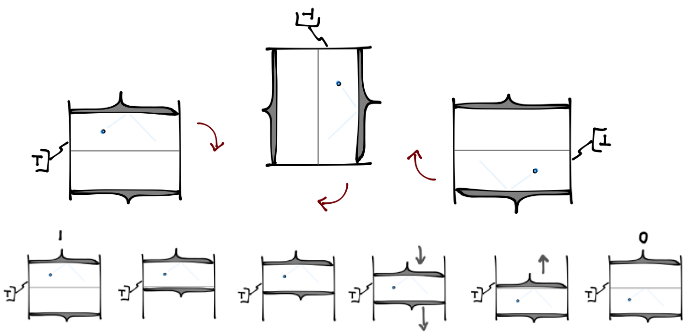
$$W = \int F dx = \int P A dx = \int P dV \overset{\text{ideal gas}}{=} \underbrace{\int_{V_0}^{V_f} \frac{dV}{V}}_{=\ln 2} 1 k_B T = k_B T \ln 2 .$$

We can use this work to lift a weight. [End of Lecture 6]

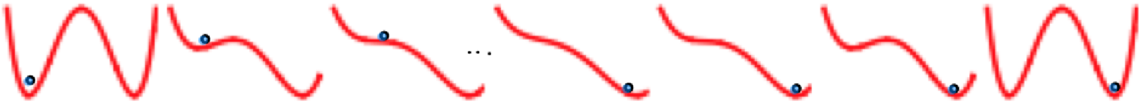
If someone hands us a memory tape with a string of *known* bits, we can use it to drive our locomotive, by doing the procedure above as each cell goes past. When the tape comes out of the locomotive, the bits are completely randomized. A uniformly random tape is useless. Only to the extent that we can predict the next bit can we do work. The entropy available to do work is then $N - H(p)$ where p is the probability distribution on the N bits of the tape.



Notice that we can reversibly convert a known 0 to a 1. This is like a NOT gate. There are two ways to do this in our realization. One is just to rotate the box! The other is easier to explain with pictures. The important thing is that no compression of the gas is involved.



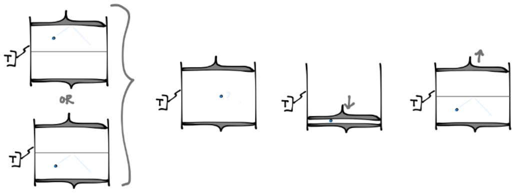
Independence of the hardware. Instead of the silly one-molecule classical ideal gas, any "bistable physical system" can serve as a one-bit memory device for the present discussion. What does this phrase mean? It means a system that is described by a double-well potential for some variable. For example, this could be the Landau-Ginzburg-Wilson free energy for a ferromagnet as a function of the magnetization.



0 → 1

We can also reversibly copy (classical!) information. The idea is to take another memory (in some unknown state), and adiabatically couple it to our system in such a way that it ends up in the same state. This process is depicted above. The ... is the delicate part which must be done slowly to avoid the acquisition of kinetic energy by the particle (which would have to be dissipated, making the process irreversible). It is important that the state of the register onto which we are copying is known in advance (here it was 0).

But *erasing* a bit is a problem. By erasing an unknown bit, we mean the process depicted at right:
This use of the term 'erasure' is debatable: it might be better to call it *resetting*; we are resetting the bit to a reference state. We might want to do this, for example, in order to define a cycle of a putative information engine (more below).



Notice that we don't find out what it was. This is absolutely crucial: the dissipative, irreversible, costly step is erasing an *unkown* bit. If we *know* the value of the bit, we can reset it for free (if it's 0, just leave it alone, and if it's 1 use the reversible conversion procedure above). But in that case the information *has not been erased* – it's still in our head! All we've done is throw away the copy in the gas memory!

Another crucial point is that in the copy procedure described above, we must know the initial state of the register onto which we do the copy. (We don't need to know the state of the register which is being copied.) Otherwise, this is the same as erasing the target register, and is dissipative. Here's a proof: there are two possible final states of this procedure (corresponding to the two possible states of the bit being copied), but if the initial state of the copy-to bit is unknown, there are four possible initial states. So there is no way this procedure could be reversible.

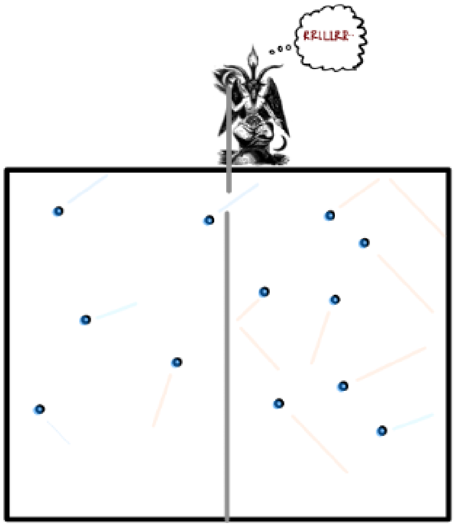
Notice that burning information as fuel and erasing the information are opposite processes.

Landauer's principle: *Erasure of information is invariably accompanied by the generation of heat.* The dissipation is associated with the logical irreversibility of the operation.

Like many thermodynamic arguments, this statement can be demonstrated by showing it in some particular realization (like a steam engine) and then using the fungibility of energy (*i.e.* our ability to convert energy between various systems) to argue that it must hold in any realization. Here we must also appeal to the fungibility of information.

Exercise: In the realization of a bit as a one-molecule gas, it is clear that resetting an unknown bit to a reference state (say 0) requires energy at least $kT \ln 2$. In the realization with the general double-well potential, how do we see that we can't just use the copy procedure on an *unknown* bit to set it for free equal to a reference value? Bennett gives an answer on page 933.

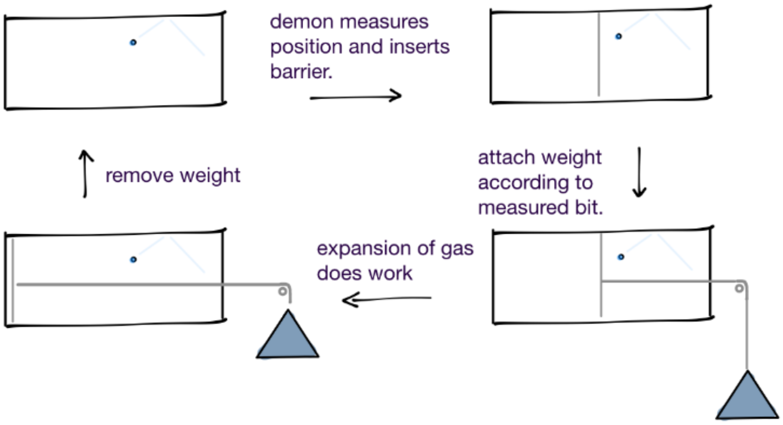
Maxwell demon. Historically the first version of this discussion is due to Maxwell, a very smart person. If you need some humility in your life, consider that Maxwell lived before the existence of atoms was widely accepted.
Imagine a box of gas divided into two halves. A demon sits at an aperture in the partition and lets the fast molecules go through to the right and the slow molecules go through to the left. In this way the demon can generate a temperature gradient which can be used to do work.



The same principle can be used to create an apparent violation of the second law in the form

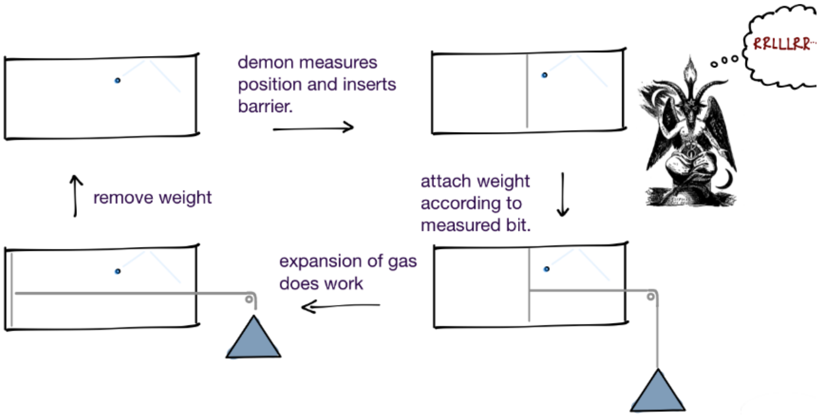
A cycle of a closed system cannot have as its only result the conversion of heat to work.

This is called a *Szilard engine.*


demon measures position and inserts barrier.
attach weight according to measured bit.
expansion of gas does work
remove weight

The net effect of the cycle depicted above right seems to be to extract work from the heat bath, period. For a long time it was believed that it was the process of measurement that was the difficulty. But it is not. The difficulty is that it is not in fact a cycle of a closed system: we have left out the state of the demon.[19] We can model the demon's memory by another bistable physical system; classically, measurement just means copying the state of the system into the (initialized!) demon memory. We argued above that this can be done reversibly, as long as the state of the demon's memory is initialized.

However, this realization that the demon is a physical system shows where the problem is: the demon stores the information in some physical system which acts as a memory. To use it again, the memory must be reset. It is governed by physics!


demon measures position and inserts barrier.
attach weight according to measured bit.
expansion of gas does work
remove weight
RRLLLRR...

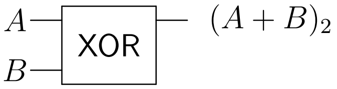
The finiteness of the demon's memory saves the 2d Law of Thermodynamics. The simplest model of the demon's memory is just a two-state system; to make a cycle we would need to erase (*i.e.* reset, or initialize) the bit. This costs

$$W_{\text{Landauer}} \geq -k_B T \ln 2$$

which is transferred as heat (say during the weight-removal step) back to the reservoir $\Delta Q = T\Delta S_{\text{system}}$. The net result is nothing happens, at best.

Reversible computation. One important scientific outcome of this line of work (by Maxwell, Szilard, Feynman, Landauer, Bennett) is the realization that computation can be reversible, and there is no minimum energy cost.

Consider an XOR gate:



Here's a specious argument that this process cannot be done reversibly (in italics because it is wrong): *The output is zero or one. Whichever outcome obtains compresses the phase space by a factor of two. Therefore* $F \geq kT \ln 2$ *is required.*

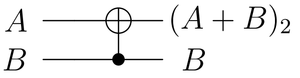
[19] Amusingly, the confusions associated with both the Maxwell demon and the Schrödinger cat arise from failing to include the observer(s) as part of the physical system.

A more important and correct point is that we cannot reconstruct the input from the output. The operation cannot be undone, because there is not enough information to reverse it. But surely this can be done reversibly:

$$A - \boxed{\mathsf{NOT}} - \ (A+1)_2$$

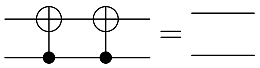
Here we just flip the bit. If we do this twice, we do nothing: $- \boxed{\mathsf{NOT}} - \boxed{\mathsf{NOT}} - = -$.

Now consider instead a gate which takes two bits as input and outputs *two* bits. One of the outputs is just the same as the XOR gate output, and other is just one of the inputs:



This is called CNOT or controlled-NOT or controlled-X or CX.

If we do it twice we do nothing: it is invertible, in particular it's its own inverse:



$$\mathsf{CX}^2 = \mathbb{1}.$$

The only inescapable energy cost comes at the step when we take out the garbage to reset the device.

This realization played an important role in leading people to think about quantum computers. [Benioff] I believe that (at least from a certain very abstract viewpoint) reversible computation just means quantum computation without entanglement or superposition.

Here's what I mean. Regard the bits A, B above as qbits which happen to be eigenstates of $\mathbf{Z}$ (recall that this means σ^z), and we call the eigenstates $|\uparrow = 0\rangle, |\downarrow = 1\rangle$. (Note that $s = 0, 1$ are the eigenvalues of $\frac{\log \mathbf{Z}}{\mathbf{i}\pi}$, in the sense that $\mathbf{Z}|s\rangle = e^{\mathbf{i}\pi s}|s\rangle$. Alternatively, $\mathbf{s} = \frac{1}{2}(1 + \mathbf{Z})$ is the projector onto states with spin up and s is its eigenvalue.) The NOT operator is then just $\mathbf{X}$:

$$\mathbf{X}|s\rangle = |(s+1)_2\rangle .$$

And the operator control-X can be written variously as

$$\mathsf{CX}_{BA} = |0\rangle\langle 0|_B \otimes \mathbb{1}_A + |1\rangle\langle 1|_B \otimes \mathbf{X}_A = \mathbf{X}_A^{\frac{1}{2}(1-\mathbf{Z}_B)} = e^{\frac{\mathbf{i}\pi}{4}(1-\mathbf{Z}_B)(1+\mathbf{X}_A)}.$$

Notice that $\mathbf{X}_A$ and $\mathbf{Z}_B$ commute so I didn't need to worry about operator ordering in the above gymnastics.

From this point of view, it is clear how to do reversible computations: only use unitary gates.

Some comments:

- According to Feynman (*Computation*, section 5) and Plenio-Vitelli, the Landauer principle can be used to motivate the Shannon noisy channel theorem, but I haven't understood this discussion. Let me know if you do.
- Some of the reversible operations above required us to do things arbitrarily slowly. You might worry about this tradeoff between reversibility and finite computation speed. Feynman section 5.3 makes some estimates of the free energy cost of doing things at a finite rate. If we work in thermal equilibrium at temperature T, and the two states between which our computation runs have energies $E_1 > E_2$, we might expect the rate to be proportional to the Boltzmann factor $r \propto e^{-\beta(E_1 - E_2)}$. Solving this equation for the energy difference suggests
$$\Delta E \sim k_B T \log r.$$
It is not clear to me whether this can be regarded as a lower bound for a given rate.
- Biomolecules do this kind of 'Brownian computation' which can happen reversibly in either direction, but is pushed in one direction by some osmotic pressure from the availability of reactants. For more on this, see Sethna chapter 5, Feynman 5.2. A more complete discussion of the kind of polymer synthesis and copying they are talking about should mention *kinetic proofreading*, for which see *e.g.* Bialek's biophysics textbook.
- Here is an attempt to give some rigorous underpinnings to Landauer's principle.

3.2 Second Laws of Thermodynamics

[C&T, chapter 4; MacKay, chapter 8]

I would like to spend a little bit of time thinking about results in information theory which resemble the Second Law of Thermodynamics. Generally, the goal is to identify irreversibility.

Define a *stochastic process* as a collection $\{X_1 \cdots X_N\}$ of random variables indexed by a variable $n = 1...N$ which we'll regard as time. They are *not* necessarily independent. Such a process is called stationary if the joint distribution for all subsets is invariant under a time shift, $n \to n+1$. Stationary distributions determine the possible long-term behavior, $n \to \infty$.

A process is a *Markov process* if its memory does not last beyond one time step, *i.e.*

$$p(X_{n+1}|X_n \cdots X_1) \overset{\text{Markov}}{=} p(X_{n+1}|X_n).$$

This means that the joint distribution can be written as

$$p(X_1 \cdots X_n) = p(X_n|X_{n-1})p(X_{n-1}|X_{n-2}) \cdots p(X_2|X_1)p(X_1).$$

And the distribution for the next time in terms of the current one is

$$p(X_{n+1}) = \sum_{x_n} \underbrace{p(X_{n+1}|X_n = x_n)}_{\equiv P_{n+1,n}} p(x_n).$$

The quantity P is a transition matrix. So a Markov process is just concatenated noisy channels:

$$X_1 \to \boxed{p(X_2|X_1)} \to X_2 \to \boxed{p(X_3|X_2)} \to X_3 \to \boxed{p(X_4|X_3)} \to ...$$

The statement that $X_1X_2X_3$ form a Markov chain is therefore abbreviated as $X_1 \to X_2 \to X_3$ (omit the boxes).

A stationary Markov distribution means both that $p(X_{n+1}|X_n) = P$ is independent of n, and that the input to each P is reproduced by P: $\mu_j = \sum_i \mu_i P_{ij}, \forall j$. (I say *a* stationary distribution because there could be more than one basin of attraction.)

In terms of these notions we can state various facts which govern the time dependence of the entropy, like the second law of thermodynamics does.

(1) Let μ_n, μ'_n be two families of distributions resulting from the *same* Markov process. Their relative entropy $D(\mu_n||\mu'_n) \equiv \delta_n$ *decreases* with n, *i.e.* $\delta_n \geq \delta_{n+1}$.

Consider the joint distribution for two successive steps:

$$p(x_n, x_{n-1}) = p(x_{n+1}|x_n)p(x_n)$$

and the same for primes:

$$p'(x_n, x_{n-1}) = p(x_{n+1}|x_n)p'(x_n)$$

(note that there is no prime on the transition matrix, since they are evolving by the same Markov process). Let $\mu_n \equiv p(X_n)$ be the nth marginal.

(Lemma:) The relative entropy for a joint distribution satisfies a chain rule in the form

$$D(p_{xy}||q_{xy}) = D(p_x||q_x) + D(p(y|x)||q(y|x)). \tag{3.1}$$

Here $D(p(y|x)||q(y|x)) \equiv \sum_x p_x \sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)}$ is (implicitly) defined to be the average over $p(x)$. Like the chain rule for entropy, (3.1) follows from the definitions and a liberal use of Bayes equation (see page 25 of C&T for a proof which leaves nothing to the imagination). The same equation holds with the roles of x and y switched.

Apply both of these to the joint distribution for two successive steps:

$$\begin{aligned} D\left(p(x_n, x_{n+1}||p'(x_n, x_{n+1})\right) = & \quad \underbrace{D(p(x_n)||p'(x_n))}_{=\delta_n} + \underbrace{D(p(x_{n+1}|x_n)||p'(x_{n+1}|x_n))}_{=0\text{, since the two distr. are the same}} \\ = & \quad \underbrace{D(p(x_{n+1})||p'(x_{n+1}))}_{=\delta_{n+1}} + \underbrace{D(p(x_n|x_{n+1})||p'(x_n|x_{n+1})}_{\geq 0} \end{aligned} \tag{3.2}$$

The equation in the underbraces is the one we are after. ■

So: using the relative entropy as a measure of distance, the Markov evolution from any two initial conditions produces more and more similar distributions – as if they were converging to some equilibrium distribution. Indeed:

(2) Apply the first equation in (3.2) with $\mu' = \mu^\star$ chosen to be any stationary distribution for the process in equation, *i.e.* $\mu^\star_n = \mu^\star_{n+1}$. So

$$D(\mu_n||\mu^\star) \geq D(\mu_{n+1}||\mu^\star)$$

– μ_n gets closer to any stationary distribution as time goes on. Such a monotonically non-increasing *positive* sequence as these δ_ns has a limit, and that limit is zero if $\mu^\star$ is unique.

(3) You may notice something awkward about the above: the 2d law is usually stated in some form involving the words "entropy increases over time", which seems semantically opposite of what we've just said.

But indeed, IFF the uniform distribution $u(x) \equiv \frac{1}{|\mathcal{X}|}$ (recall that $|\mathcal{X}|$ is the number of elements of the sample set) is stationary, then

$$H\left(\mu_n\right) \leq H\left(\mu_{n+1}\right),$$

the Shannon entropy increases.

$$\boxed{\Longrightarrow:} \qquad \underbrace{D(\mu_n||u)}_{\text{shrinks with } n} = \sum_x \mu_n(x)\log\left(\frac{\mu_n(x)}{u}\right) = \underbrace{\log|\mathcal{X}|}_{\text{ind. of } n} - \underbrace{H(\mu_n)}_{\text{grows with } n} \quad \Longrightarrow$$

$\boxed{\Longleftarrow:}$ If the uniform distribution $u(x) = \frac{1}{|\mathcal{X}|}$ is *not* stationary, it evolves to a stationary one $\mu^\star$ (by result (2) above). But the uniform distribution is the maximum-entropy distribution on this set (since $\forall p$,

$$0 \leq D(p(x)||u) = \log|\mathcal{X}| - H(p)$$

and equality only holds if $p = u$) so in this case

$$H(u) = \log |\mathcal{X}| > H(\mu_\star)$$

and we've shown that $H(\mu_n) \leq H(\mu_{n+1})$ doesn't hold if u isn't stationary. ■

This begs the question: under what circumstances is the uniform distribution stationary, $u = \mu^\star$?
Claim: u is stationary IFF

$$P_{ij} \equiv p(i|j) \equiv \text{Prob}(x_n = j|x_{n-1} = i)$$

is *doubly stochastic* which means P^t is also a probability distribution, $\sum_i P_{ij} = 1, \ \forall j$. (In particular this holds if $P_{ij} = P_{ji}$ is symmetric.)

Instructions for proof: stare at the condition that u is stationary $Pu = u$.

Unproved claim: A doubly stochastic distribution is a convex combination of permutations (a permutation is a transition matrix with just one nonzero entry in each row and column).

(4) **Data-processing inequality.** [MacKay problem 8.5, Shumacher §20.1

[End of Lecture 7]

Consider a Markov chain $p(XYZ) = p(Z|Y)p(Y|X)p(X)$ which relationship we can denote $X \to Y \to Z$. In words: if we know Y for sure, we don't learn more about Z from learning X. More elegantly: the associated conditional mutual information vanishes

$$I(Z : X|Y) = 0.$$

(Recall that $I(Z : X|Y) \equiv D(p(ZX|Y)||p(Z|Y)p(X|Y)) = \left\langle \log \frac{p(ZX|Y)}{p(Z|Y)p(X|Y)} \right\rangle_{XYZ} = H(Z|Y) - H(Z|YX)$.) In fact, since the relative entropy only vanishes for equality, this vanishing of the conditional mutual info is equivalent to the Markov property. And since $I(Z : X|Y) = I(X : Z|Y)$ is symmetric in Z, X, this means that $Z \to Y \to X$ is also a Markov chain.

The data-processing inequality is

$$X \to Y \to Z \quad \implies \quad I(X : Y) \geq I(X : Z).$$

The proof follows by the same trick (as in (3.2)) of using the chain rule twice. The mutual information satisfies the following chain rule (proved by the same methods as the others):

$$I(X : YZ) = I(X : Z) + \underbrace{I(X : Y|Z)}_{\geq 0} = I(X : Y) + \underbrace{I(X : Z|Y)}_{\overset{\text{Markov}}{=} 0} \tag{3.3}$$

■

Equality holds IFF $X \to Z \to Y$ also.

Another related fact is

$$I(X:Y|Z) \overset{\text{Markov},(3.3)}{=} I(X:Y) - \underbrace{I(X:Z)}_{\geq 0} \leq I(X:Y)$$

which says observing Z can't decrease the dependence of X and Y. (We saw examples where it could increase it. This means that in the case where $X \to Y \to Z$ are Markov, the area of that middle region in the Venn diagram near (2.9) is actually always ≥ 0.)

Notice that $X \to Y \to f(Y)$ is Markov, where f is some deterministic operation. For example: suppose we have a noisy channel $p(Y|X)$, X is the sent message and Y is the received message. Let $f(Y)$ be the receiver's estimated decoding of the message. Clearly this is a Markov process because $f(Y)$ only knows about Y and not X (otherwise we don't need to estimate).

From this point of view, the data-processing theorem says that processing (doing operations $f(Y)$) can only destroy information.

4 Quantifying quantum information and quantum ignorance

4.1 von Neumann entropy

[A good source is: Schumacher §19.3] A density matrix $\boldsymbol{\rho}$ acting on $\mathcal{H}$ is a generalization of a probability distribution. Our job here is to understand and make precise this statement. In this discussion we can be agnostic about the origin of the density matrix: it could be that someone is shooting an electron gun whose output comes from some ensemble $p(X)$ of set of (not necessarily orthogonal) quantum states $|\psi_x\rangle$ (in which case $\boldsymbol{\rho} = \sum_x p(x)|\psi_x\rangle\langle\psi_x|$), or perhaps $\mathcal{H}$ is a subspace of a larger Hilbert space to which we do not have access. Each density matrix can be constructed in many ways.

Inherent in a density matrix are two sources of uncertainty: uncertainty about which is the quantum state, and quantum uncertainty of measurements of non-diagonal operators.

One thing about which we are sure is that the density matrix is positive semi-definite (hence hermitian) and has $\text{tr}\boldsymbol{\rho} = 1$. Its hermiticity guarantees a spectral decomposition

$$\boldsymbol{\rho} = \sum_a p_a \left|a\right\rangle \left\langle a\right|,$$

and the other properties guarantee that the p_a are probabilities: $p_a \in [0,1], \sum_a p_a = 1$. They may be interpreted as the probability that the quantum state is (the $\boldsymbol{\rho}$-eigenstate) $|a\rangle$.

Functions of a hermitian operator can be defined in terms of the spectral decomposition: $f(\boldsymbol{\rho}) \equiv \sum_a f(p_a) \left|a\right\rangle \left\langle a\right|$, so in particular $\log\boldsymbol{\rho} = \sum_a \log(p_a) \left|a\right\rangle \left\langle a\right|$ and even better (since there is no trouble with $p_a = 0$ in this case)

$$-\boldsymbol{\rho}\log\boldsymbol{\rho} = -\sum_a p_a \log(p_a) \left|a\right\rangle \left\langle a\right|$$

is a hermitian operator on $\mathcal{H}$ and its trace is

$$S(\boldsymbol{\rho}) \equiv -\text{tr}\boldsymbol{\rho}\log\boldsymbol{\rho} = -\sum_a p_a \log(p_a) = H(p),$$

the von Neumann entropy of $\boldsymbol{\rho}$. It is a basis-independent functional of $\boldsymbol{\rho}$. In the specific context in which $\boldsymbol{\rho}$ is a reduced density matrix arising by tracing out some part of a larger Hilbert space, this is also called the *entanglement entropy*. Let us consider its qualities as a measure of the quantum information contained in $\boldsymbol{\rho}$, by analogy with the Shannon entropy.

To get started, you may say: no big deal, it is just the Shannon entropy of the set of eigenvalues. But consider the following. We showed that the Shannon entropy for a joint distribution satisfies the perhaps-intuitive property that $H(XY) \geq H(Y)$ – the entropy of the whole is bigger than the entropy of a part.[20] The quantum analog of a joint distribution is a bipartite state $\boldsymbol{\rho}_{AB}$ on $\mathcal{H}_A \otimes \mathcal{H}_B$. Consider for example the case when both $\mathcal{H}_{A,B}$ are qbits, and we take a pure state

$$\boldsymbol{\rho}_{AB} = |\text{Bell}\rangle \langle\text{Bell}|\,, \quad |\text{Bell}\rangle \equiv \frac{|\downarrow\uparrow\rangle - |\uparrow\downarrow\rangle}{\sqrt{2}}.$$

Now, for any pure state (by definition a density matrix which is a rank-one projector $\boldsymbol{\rho}^2_{\text{pure}} = \boldsymbol{\rho}_{\text{pure}}$) there is only one nonzero eigenvalue (which must be one) $S(\rho_{\text{pure}}) = H(\{1, 0..0\}) = 0$, and in particular, the 'quantum entropy of the whole' in this case is zero.

What's the 'quantum entropy of part'? We must find $S(\boldsymbol{\rho}_A)$ with

$$\boldsymbol{\rho}_A = \text{tr}\, |\text{Bell}\rangle \langle\text{Bell}|\,.$$

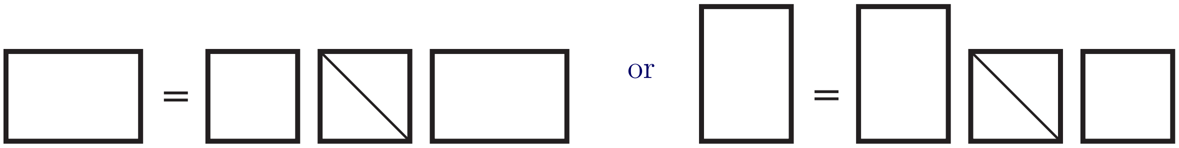
In this case, we can do it by hand and the answer is $\boldsymbol{\rho}_A = \frac{1}{2}\mathbb{1}$, whose entropy is $S\left(\frac{1}{2}\mathbb{1}\right) = 1$. Quantumly, the entropy of the parts can be larger!

Why you should love the Schmidt decomposition. More generally, it will be useful to discuss the notion of Schmidt decomposition of a bipartite state $|w\rangle = \sum_{aj} w_a^j \,|a\rangle_A \,|j\rangle_B$. The singular value decomposition (SVD) of the matrix w is

$$w = UsV, \quad i.e. \quad w_a^j = \sum_{r=1}^{\chi} U_a^r s_r V_r^j \tag{4.1}$$

where s_r are the singular values, and if we want to keep the einstein summation convention, we should write s as a diagonal matrix. U and V are unitary, and χ is the Schmidt rank. Depending on whether A or B is bigger, the SVD (4.1) looks like (left and right respectively):

[20] This follows from the fact that $0 \leq H(X|Y) = H(XY) - H(Y)$. The positivity follows since $H(X|Y) = \langle p(X|Y = y)\rangle_Y$ is an average of Shannon entropies (each positive). The novelty quantum mechanically is that there is no well-defined notion of conditional probability! The quantity $S(XY) - S(Y)$ makes perfect sense and we can call it $S(X|Y)$ 'the conditional entropy' but it is *not* an average of any kind of 'conditional von Neumann entropies', and indeed it can be negative. Note that the difficulty of defining such conditional entropies in quantum mechanics underlies many of the deepest facts.



[Figure from U. Schöllwock, *DMRG in the age of MPS*]. The unitaries U and V can be used to define a partial basis for A, B, so that we may write $|r\rangle_A \equiv U^r_a |a\rangle$, $|r\rangle_b \equiv V^j_r |j\rangle_B$, and

$$|w\rangle = \sum_{r=1}^{\chi} s_r |r\rangle_A \otimes |r\rangle_B .$$

This is the Schmidt decomposition. The unitary property of U, V guarantee that the states $\{|r\rangle_A\}, \{|r\rangle_B\}$ are each orthonormal (though the ones in the larger space will not be complete)[21], so that *e.g.* $\langle r|r'\rangle_B = \delta_{r,r'}$. Here's the payoff:

$$\boldsymbol{\rho}_A = \mathrm{tr}_B |w\rangle\langle w| = \sum_{r=1..|B|} \sum_{r_1,r_2} {}_B\langle r| \left(|r_1\rangle_B \otimes |r_1\rangle_A s_{r_1} s^\star_{r_2} \langle r_2|_A \otimes \langle r_2|_B\right) |r\rangle_B = \sum_{r=1}^{\chi} s_r s^\star_r |r\rangle_A \langle r|_A$$

The eigenvalues of $\boldsymbol{\rho}_A$ are $p_r = |s_r|^2$. When $\boldsymbol{\rho}$ comes from tracing out part of a larger system, the logs of the eigenvalues of p_r are sometimes called the *entanglement spectrum*.

Notice that these are also the eigenvalues of $\boldsymbol{\rho}_B = \mathrm{tr}_A |w\rangle\langle w|$. If the whole system is in a pure state, the vN entropy (and indeed the whole entanglement spectrum) of A and its complement $\bar{A}$ are equal.

The largest the vN entropy can be is $S(u = \mathbb{1}/|\mathcal{H}|) = \log|\mathcal{H}|$; if the system is a collection of qbits, $\mathcal{H} = \otimes_x \mathcal{H}_x$, this is just the number of qbits. (And if they are extended in space, this is proportional to the *volume* of space.) We can prove this by the same method as we used for Shannon: the relative entropy. Wait for it – §4.2.

'Sampling a density matrix'. To continue pushing the analogy with classical probability distributions, what does it mean to sample a density matrix $\boldsymbol{\rho}$ with spectral decomposition $\boldsymbol{\rho} = \sum_k \rho_k |k\rangle\langle k|$ on $\mathcal{H}$? Whatever this means, it should produce a random pure state in $\mathcal{H}$. Unlike the classical case, this is not a uniquely defined procedure. In particular, (I believe) to make this well defined, we must specify an observable $\mathbf{A} = \mathbf{A}^\dagger = \sum_n a_n |a_n\rangle\langle a_n|$ on $\mathcal{H}$. $\mathbf{A}, \boldsymbol{\rho}$ together produce a classical distribution $p(A)$

[21] The way I've drawn the picture here, U and V are actually not whole unitaries (a unitary matrix must be square!), but rather *isometries*. This means $\sum_a \Upsilon^\dagger_{ra}\Upsilon_{ar'} = \mathbb{1}_{rr'}$ (like a unitary) but $\sum_r \Upsilon_{ar}\Upsilon^\dagger_{rb}$ has smaller rank because there aren't enough terms in the sum over r to resolve the identity. Note by the way that if Υ is an isometry, then $\Upsilon^\dagger$ is called a partial isometry. If we instead define the matrix s to be rectangular, by filling in the rest with zeros, $s^{r'}_r = 0, r, r' = \chi ... \max|A|, |B|$, then we can let U, V be unitary. Thanks to Sami Ortoleva for reminding me that this is a better convention.

for a random variable $a \in \{a_n\}$ (the outcome of a measurement of $\mathbf{A}$) with

$$p(a_n) \equiv \text{Prob}(A = a_n) = \text{tr}\boldsymbol{\rho} \left|a_n\right\rangle \left\langle a_n\right| = \sum_k |\left\langle a_n|k\right\rangle|^2 \rho_k \equiv \sum_k M_{nk}\rho_k.$$

(In the penultimate step I assumed the eigenvalues of $\mathbf{A}$ were nondegenerate for simplicity.) The random state resulting from sampling (in this sense) is then $|a_n\rangle$ with probability $p(a_n)$.

(Note that the matrix $M_{nk} \equiv |\left\langle a_n|k\right\rangle|^2 \geq 0$ is doubly stochastic: $\sum_n M_{nk} = 1, \forall k, \sum_k M_{nk} = 1, \forall n$; it is a probability distribution on both arguments.)

Now we can consider the Shannon entropy of the RV $p(A)$:

$$\begin{aligned}
H(A) &= -\sum_n p(a_n)\log p(a_n) \\
&= -\sum_n \left(\sum_k M_{nk}\rho_k\right) \log \left(\sum_{k'} M_{nk'}\rho_{k'}\right) \\
&\overset{f(x)\equiv -x\log x,\ \langle\rho\rangle_n \equiv \sum_k M_{nk}\rho_k}{=} \sum_n f(\langle\rho\rangle_n) \overset{f(\langle R\rangle)\geq\langle f(R)\rangle}{\geq} -\sum_n\sum_k M_{nk}\rho_k \log\rho_k \\
&\overset{\sum_n M_{nk}=1}{=} S(\boldsymbol{\rho}) \ . \qquad (4.2)
\end{aligned}$$

The preceding seems forbidding but the conclusion is unsurprising if we recall the extra quantum uncertainty: even if we know the quantum state, *e.g.* of a single qbit, for sure, $\boldsymbol{\rho} = |0\rangle\langle 0|$, measuring a non-eigenstate (*e.g.* $\mathbf{A} = \mathbf{X}$), the outcome is uncertain.

4.2 Quantum relative entropy

Given $\boldsymbol{\rho}, \boldsymbol{\sigma}$ density matrices on $\mathcal{H}$, the quantum relative entropy is

$$\hat{D}(\boldsymbol{\rho}||\boldsymbol{\sigma}) \equiv \text{tr}\boldsymbol{\rho}\log\boldsymbol{\rho} - \text{tr}\boldsymbol{\rho}\log\boldsymbol{\sigma}.$$

I will sometimes put a hat on it to distinguish it from the classical relative entropy.

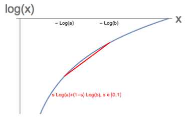
Fact:

$$\hat{D}(\boldsymbol{\rho}||\boldsymbol{\sigma}) \geq 0, \ \forall\boldsymbol{\rho}, \boldsymbol{\sigma}.$$

Proof: let their spectral representations be $\boldsymbol{\rho} = \sum_k \rho_k |k\rangle\langle k|, \boldsymbol{\sigma} = \sum_n \sigma_n |s_n\rangle\langle s_n|$ and recall $\log\boldsymbol{\sigma} = \sum_n |s_n\rangle\langle s_n|\log\sigma_n$. Then

$$\begin{aligned}\hat{D}(\boldsymbol{\rho}||\boldsymbol{\sigma}) &= \sum_k \rho_k \log \rho_k - \sum_k \rho_k \sum_n \underbrace{\langle k|s_n\rangle \langle s_n|k\rangle}_{=M_{nk}} \log \sigma_n \\ &\geq \sum_k \rho_k \left(\log \rho_k - \log \tau_k\right) \\ &= \sum_k \rho_k \log \frac{\rho_k}{\tau_k} = D(\rho_k||\tau_k) \geq 0.\end{aligned}$$

Log is concave:



$$\implies \sum_n M_{nk} \log \sigma_n \leq \log \left(\underbrace{\sum_n M_{nk} \sigma_n}_{\equiv \tau_k} \right)$$

In this last step, this is just a classical relative entropy which we know is positive. Equality holds iff $\boldsymbol{\rho} = \boldsymbol{\sigma}$. ■[22]

Here's an immediate application of the positivity of the quantum relative entropy: its positivity means the uniform density matrix $\mathbf{u} \equiv \frac{1}{|A|}\mathbb{1}_A$ has a larger entropy than any other density matrix $\boldsymbol{\rho}$ on A:

$$0 \leq \hat{D}(\boldsymbol{\rho}||\mathbf{u}) = \mathrm{tr}_A \boldsymbol{\rho} \log \boldsymbol{\rho} - \mathrm{tr}_A \boldsymbol{\rho} \log \mathbf{u} = -S(\boldsymbol{\rho}) + \log |A| \qquad \blacksquare$$

Here's another, closely-related application: Recall that the thermal equilibrium density matrix at temperature T for a system with Hamiltonian H is

$$\boldsymbol{\rho}_T = Z^{-1} e^{-\frac{\mathbf{H}}{k_B T}}, \quad Z \equiv \mathrm{tr}_{\mathcal{H}} e^{-\frac{\mathbf{H}}{k_B T}}.$$

Its vN entropy is

$$S(\boldsymbol{\rho}_T) = \frac{\log e}{k_B T} \mathrm{tr} \mathbf{H} \boldsymbol{\rho}_T + \log Z = \frac{\log e}{k_B T} \langle \mathbf{H} \rangle_{\rho_T} + \log Z$$

which up to the overall normalization is the thermal entropy, $S = -\partial_T F = -\partial_T \left(-k_B T \ln Z\right)$.

Claim: the thermal state has the maximum entropy for any state with the same expected energy $E = \langle \mathbf{H} \rangle$. This is true since for any other $\boldsymbol{\rho}$ with $\mathrm{tr}\boldsymbol{\rho}\mathbf{H} = E$,

$$0 \leq D(\boldsymbol{\rho}||\boldsymbol{\rho}_T) = -S(\boldsymbol{\rho}) + \mathrm{tr}\boldsymbol{\rho} \log \frac{e^{-\frac{\mathbf{H}}{k_B T}}}{Z}$$

[22] The positivity of the quantum relative entropy is a special case of *Klein's inequality*, which is: for any two positive linear operators on $\mathcal{H}$, $\mathbf{A}, \mathbf{B} > 0$,

$$\mathrm{tr}_{\mathcal{H}} \mathbf{A} \left(\log \mathbf{A} - \log \mathbf{B}\right) \geq \mathrm{tr}_{\mathcal{H}} (\mathbf{A} - \mathbf{B})$$

with equality iff $\mathbf{A} = \mathbf{B}$. This more general version will be useful in proving strong subadditivity. It can be seen to be equivalent to the version we proved above by writing $\boldsymbol{\rho} \equiv \mathbf{A}/\mathrm{tr}\mathbf{A}, \boldsymbol{\sigma} \equiv \mathbf{B}/\mathrm{tr}\mathbf{B}$ and using $\log x \leq x - 1$. This in turn is a special case of the following identity (also named after Klein I think, and which I learned about from Wehrl) which says that for any convex function $f(x)$ and pair of positive linear operators,

$$\mathrm{tr}\left(f(\mathbf{B}) - f(\mathbf{A})\right) \geq \mathrm{tr}\left(\mathbf{B} - \mathbf{A}\right) f'(\mathbf{A}).$$

The previous version obtains when $f(x) = -x \log x$.

$$= -S(\boldsymbol{\rho}) + \frac{\log e}{k_B T} E + \log Z = -S(\boldsymbol{\rho}) + S(\boldsymbol{\rho}_T) \ . \tag{4.3}$$

This is a step towards a Bayesian point of view on why we should use the canonical density matrix in the first place. [End of Lecture 8]

(Quantum) mutual information. Given $\boldsymbol{\rho}_{AB}$ on a bipartite $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$,

$$S(A:B) \equiv \hat{D}\left(\boldsymbol{\rho}_{AB} || \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B\right)$$

where $\boldsymbol{\rho}_A \equiv \mathrm{tr}_B \boldsymbol{\rho}_{AB}$ and $\boldsymbol{\rho}_B \equiv \mathrm{tr}_A \boldsymbol{\rho}_{AB}$ are the partial traces (the analog of marginals).

In terms of vN entropies, it is (just like in the classical case)

$$S(A:B) = S(A) + S(B) - S(AB).$$

And since it is a relative entropy, it is positive: $S(A:B) \geq 0$, which implies *subadditivity* of the vN entropy: $S(A) + S(B) \geq S(AB)$.

4.3 Purification, part 1

Here is a beautiful idea due to Araki and Lieb, I believe. Given $\boldsymbol{\rho}_{AB}$ on a bipartite $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, the vN entropies participate in the following 'triangle inequality'

$$|S(\boldsymbol{\rho}_A) - S(\boldsymbol{\rho}_B)| \leq S(\boldsymbol{\rho}_{AB}).$$

The idea of the proof is to introduce an auxiliary system C which *purifies* the state $\boldsymbol{\rho}_{AB}$:

$$|\psi\rangle \in \mathcal{H}_{ABC} \quad \text{with} \quad \mathrm{tr}_C \, |\psi\rangle \langle\psi| = \boldsymbol{\rho}_{AB}.$$

The mere existence of such a pure state then implies many statements about the entanglement entropies[23] :

$$S(C) = S(AB), \ S(AC) = S(B)...$$

by which we can eliminate the dependence on C. In particular, subadditivity on AC implies

$$S(A) + \underbrace{S(C)}_{=S(AB)} \geq \underbrace{S(AC)}_{=S(B)}$$

which says $S(B) - S(A) \leq S(AB)$. Interchanging the roles of A and B gives $S(A) - S(B) \leq S(AB)$.

[23] Note that just like for random variables, to minimize clutter, the choice of density matrix is sometimes left implicit in the expression for the entropy: $S(C) \equiv S(\boldsymbol{\rho}_C)$ etc...

• Purifications exist: If the spectral representation of $\boldsymbol{\rho} = \sum_{a=1}^{\chi_\rho} p_a |a\rangle\langle a|$ then choosing $|C| \geq \chi_\rho$, the Schmidt rank of $\boldsymbol{\rho}$, we can take

$$|\psi\rangle = \sum_a \sqrt{p_a}\,|a\rangle \otimes |a\rangle_C = \sqrt{\boldsymbol{\rho}} \otimes 1\!\!1_C \sum_{a=1}^{\chi} |aa\rangle \ .$$

This is certainly not unique: we had to make a choice of χ_ρ ON states in $\mathcal{H}_C$; any unitary rotation $\mathbf{U}_C$ of $\mathcal{H}_C$ produces another purification:

$$|\psi\rangle \mapsto (1\!\!1_{\mathcal{H}} \otimes \mathbf{U}_C)\,|\psi\rangle = \sum_a \sqrt{p_a}\,|a\rangle \otimes \mathbf{U}_C\,|a\rangle_C\,.$$

• All purifications are equivalent in the following sense: given two purifications $|\psi\rangle \in \mathcal{H} \otimes \mathcal{H}_C, |\psi'\rangle \in \mathcal{H} \otimes \mathcal{H}_D$ then $\exists$ an isometry[24] (or partial isometry, depending on which of C or D is bigger) $W : \mathcal{H}_C \to \mathcal{H}_D$ such that $(1\!\!1_{\mathcal{H}} \otimes W)\,|\psi\rangle = |\psi'\rangle$. To see this, just write the Schmidt representation of both states

$$|\psi\rangle = \sum_a \alpha_a\,|a\rangle \otimes |c_a\rangle_C\,, \quad |\psi'\rangle = \sum_a \beta_a\,|a\rangle \otimes |d_a\rangle_D \ .$$

The condition that these both purify the same state on $\mathcal{H}$ gives $p_a = |\alpha_a|^2 = |\beta_a|^2$, so the required W is just

$$W = \sum_a |d_a\rangle_D\,\langle c_a|_C\,.$$

Thermal double. An example of a purification which one encounters in various subfields of physics (such as finite-temperature quantum field theory) is a purification of the canonical density matrix

$$\boldsymbol{\rho}_T = Z^{-1} e^{-\beta \mathbf{H}} = \sum_a \frac{e^{-\beta E_a}}{Z}\,|a\rangle\langle a|$$

(the spectral decomposition of which is into energy eigenstates, and $\beta \equiv \frac{1}{k_B T}$). It is called the thermal double (or sometimes 'thermofield double'), and lives in two copies of the system Hilbert space:

$$\mathcal{H} \otimes \mathcal{H} \ni |\sqrt{\rho_T}\rangle \equiv \sum_a \sqrt{\frac{e^{-\beta E_a}}{Z}}\,|a\rangle_1 \otimes |a\rangle_2\,, \quad \mathrm{tr}_2\,|\sqrt{\rho_T}\rangle\langle\sqrt{\rho_T}| = \boldsymbol{\rho}_T.$$

---

[24] An isometry is a slice of a unitary. We defined them in footnote 21.

4.4 Schumacher compression

[Schumacher, §19.4] There is a nice quantum analog of Shannon's source coding theorem which gives an operational interpretation to $S(\boldsymbol{\rho})$ (just as Shannon's theorem gives to $H(p)$). Again it relies on a notion of (joint) typicality.

Consider repeated use of an electron dispenser: each object is associated with a Hilbert space $\mathcal{H}_Q$, and they are independently spat out in the state $\boldsymbol{\rho}$ (and never interact with each other). So the whole Hilbert space for n of them is $\mathcal{H}_{\vec{Q}} \equiv \otimes_{i=1}^n \mathcal{H}_{Q_i} \equiv \mathcal{H}_Q^n$, and the state is

$$\rho^{\vec{Q}} = \underbrace{\boldsymbol{\rho} \otimes \boldsymbol{\rho} \otimes \cdots \otimes \boldsymbol{\rho}}_{n \text{ times}} \equiv \boldsymbol{\rho}^{\otimes n}.$$

The spectral decomposition of each $\boldsymbol{\rho} = \sum_x p_x \left|x\right\rangle \left\langle x\right|$ then gives

$$\boldsymbol{\rho}^{\vec{Q}} = \sum_{x_1 \cdots x_n} \underbrace{p(x_1, \cdots, x_n)}_{=p(x_1)p(x_2)\cdots p(x_n)} \left|x_1 \cdots x_n\right\rangle \left\langle x_1 \cdots x_n\right| .$$

So we can regard the full output of the n-body dispenser as producing *sequences of $\boldsymbol{\rho}^{\vec{Q}}$ eigenstates*, labelled $X = x_1 \cdots x_n$, with probability $p(X)$, $p(x_1 \cdots x_n) = \prod_i p(x_i)$. From this set-up, we see immediately that we can apply Shannon's result in the following way:

We know from Shannon that there exists a *typical set* T of $\{x_1 \cdots x_n\}$ which contains most of the support of the distribution $p(X)$: For any given δ, ϵ, we can find T such that

$$\text{Prob}((x_1 \cdots x_n) \in T) > 1 - \delta$$

and the number of elements

$$|T| < 2^{n(H(X)+\epsilon)}$$

where $H(X)$ is the ordinary Shannon entropy of the distribution $p(X)$ (which incidentally is also $H(X) = S(\boldsymbol{\rho})$).

So far this is just the classical Shannon result. But now associated with T is a *typical subspace* $\mathcal{T} \subset \mathcal{H}_{\vec{Q}}$ with almost all the support of $\boldsymbol{\rho}$

$$\text{tr}_{\mathcal{T}} \boldsymbol{\rho}^{\vec{Q}} > 1 - \delta$$

and whose dimension is

$$\dim \mathcal{T} = |T| \leq 2^{n(S(\boldsymbol{\rho}))+\epsilon}.$$

It is sometimes useful to write

$$\text{tr}_{\mathcal{T}} ... = \text{tr}_{\mathcal{H}} \Pi ... \; ; \quad \Pi \equiv \sum_{(x_1 \cdots x_n) \in T} \left|x_1 \cdots x_n\right\rangle \left\langle x_1 \cdots x_n\right|$$

is the projector onto $\mathcal{T}$. The summary is that sampling n times from the density matrix $\boldsymbol{\rho}$ is well approximated by a uniform density matrix on the typical subspace of much smaller dimension nS.

$$\boldsymbol{\rho}^{\otimes n} \simeq 2^{-nS(\boldsymbol{\rho})}\Pi = \frac{1\!\!1_{\mathcal{T}}}{|\mathcal{T}|}.$$

So the cost per copy to store the state $\boldsymbol{\rho}$ is (asymptotically as $n \to \infty$) $S(\boldsymbol{\rho})$.

So at least, this is a useful observation when we know the density matrix $\boldsymbol{\rho}$ (for example by arduously determining it by sampling the source many times and measuring enough observables – this process, by the way, is called *state tomography*), but we want to store it in a Hilbert space C of smaller dimension $|C|$. The interesting case is when

$$S(\boldsymbol{\rho}) < |C| < |Q| \ .$$

The situation is actually better, though.

Illustration. [Barnett §8.5]: Consider, for example, the case where $\mathcal{H}_Q$ is a single qbit, and let the state be an equal-probability mixture of two states

$$\boldsymbol{\rho} = \sum_{j=0,1} \frac{1}{2} \left|\psi_j\right\rangle \left\langle\psi_j\right|$$

which, however, are not orthogonal:

$$\left|\psi_j\right\rangle = c\left|0\right\rangle - (-1)^j s\left|1\right\rangle, \quad \left\langle\psi_0|\psi_1\right\rangle = c^2 - s^2 \neq 0, \qquad c^2 + s^2 = 1.$$

So in the 'computational basis' (0, 1), $\rho = \begin{pmatrix} c^2 & 0 \\ 0 & s^2 \end{pmatrix}$ and the vN entropy of this state is $S(\boldsymbol{\rho}) = H_2(c^2)$.

Now consider n iid copies in

$$\mathcal{H}_{\vec{Q}} = \text{span}\{\left|\psi_{j_1}\right\rangle \otimes \cdots \otimes \left|\psi_{j_n}\right\rangle = \otimes_{l=1}^n \left(c\left|0\right\rangle - (-1)^{j_l}\left|1\right\rangle\right) \equiv \left|j_1 \cdots j_n\right\rangle\}$$

(Note that we are using a non-orthogonal basis here!) These basis states are equiprobable according to $\boldsymbol{\rho}^n$. How can we compress this distribution of states? A first, naive idea is to measure $\mathbf{Z} = \left|0\right\rangle\left\langle 0\right| - \left|1\right\rangle\left\langle 1\right|$ on each of them, and use the classical Shannon result. This will result, typically, in $N_0 = nc^2$ states with $j = 0$ and ns^2 states with $j = 1$. Of course, the price for knowing which are 0 is totally destroying the state we are trying to compress.

A slightly less bad idea is to measure how many zeros there are (without measuring which factors have $j = 0$). We'll get $N_0 \sim nc^2$ and after measurement the state will be

$$\left|N_0\right\rangle = (-1)^{N_0}\sqrt{\frac{N_0!(n-N_0)!}{n!}} \sum_{j_1 \cdots j_n \text{ with } N_0 \text{ zeros}} \left|j_1 \cdots j_n\right\rangle (-1)^{\sum_l j_l}$$

which is taken from only $W = \frac{N_0!(n-N_0)!}{n!} = 2^{nH_2(c^2)} \ll 2^n$ states instead of 2^n, yay.

But Schumacher's insight is that we don't actually need to measure the number of zeros, because of Shannon's source coding result: the typical states will have $N_0 = nc^2$ zeros without our doing anything about it. We can just measure the projector onto the typical subspace:

$$\Pi_T \equiv \sum_{j_1 \cdots j_n \in T} |j_1 \cdots j_n\rangle \langle j_1 \cdots j_n| \,.$$

And as long as we take $|T|$ a little bigger than $2^{S[\boldsymbol{\rho}]}$, we'll be able to reconstruct the initial state. And we can't take $|T|$ any smaller than that.

4.5 Quantum channels

For an open quantum system (such as a region of a quantum many body system, which in the below I will just call 'our subsystem A'), the laws of quantum mechanics are not the same as the ones you read about in the newspapers: the state is not a vector in $\mathcal{H}$, time evolution is not unitary, and observables aren't associated with Hermitian operators.

You understand the first statement: if our subsystem is entangled with the rest of the system, it does not have its own wavefunction, and we must use a density matrix to express our uncertainty about its quantum state. Fine.

The whole (closed) system $A\bar{A}$ evolves by unitary time evolution $|\psi\rangle_{A\bar{A}} = e^{-\mathbf{i}\int^t \mathbf{H}}\psi(0) = \mathbf{U}(t,0)\,|\psi(0)\rangle$. If the subsystem A interacts with the rest of the system A, *i.e.* $\mathbf{H}$ is *not* of the form $\mathbf{H} \overset{\text{decoupled}}{=} \mathbf{H}_A + \mathbf{H}_{\bar{A}}$, then time evolution can change the amount of entanglement between A and $\bar{A}$. How does $\boldsymbol{\rho}(t) = \text{tr}_{\bar{A}}\,|\psi(t)\rangle\langle\psi(t)|$ evolve in time? You can imagine trying to work this out by plugging in $|\psi(t)\rangle = \mathbf{U}\,|\psi(0)\rangle$, and trying to eliminate all mention of $\bar{A}$. It is useful to parametrize the possible answers. The result is another density matrix (positive, unit trace), so we know the waiting map (*i.e.* unitary waiting on the whole system followed by tracing out the environment) must be of the form

$$\boldsymbol{\rho}(0) \mapsto \boldsymbol{\rho}(t) \equiv \mathcal{E}(\boldsymbol{\rho}(0)).$$

Here $\mathcal{E}$ is a (linear) operator on operators, called a *superoperator*. Such a superoperator which specifically maps density matrices to density matrices is called a *CPTP map* or a *quantum channel*. The former stands for *completely positive and trace preserving* and just means that it respects the properties of density matrices (more anon). The latter name comes from the idea that we should think of these things as the quantum analog of a communication channel, which really means: the quantum analog of a set of conditional probabilities.

It will sometimes be useful to speak of an operator on $\mathcal{H}$ as an element of $\mathrm{End}(\mathcal{H})$ ('endomorphisms' of the vector space $\mathcal{H}$, *i.e.* homomorphisms from $\mathcal{H}$ to itself, *i.e.* linear maps on $\mathcal{H}$), and of a superoperator which takes operators on $\mathcal{H}$ to operators on $\mathcal{H}'$ as an element of $\mathrm{Hom}(\mathrm{End}(\mathcal{H}), \mathrm{End}(\mathcal{H}'))$ (short for 'homomorphisms').

To see what the possible form of $\mathcal{E}$ might look like, consider the situation where the initial state of $A\bar{A}$ is $\boldsymbol{\rho}(0)_{A\bar{A}} = \boldsymbol{\rho}_A \otimes |0\rangle\langle 0|_{\bar{A}}$ (for some reference state of the environment), and evolve by unitary time evolution

$$\boldsymbol{\rho}(0)_{A\bar{A}} \overset{\text{unitarily wait}}{\mapsto} \boldsymbol{\rho}(t)_{A\bar{A}} = \mathbf{U}\boldsymbol{\rho}(0)_{A\bar{A}}\mathbf{U}^\dagger$$

where $\mathbf{U} \sim e^{-\mathbf{i}\mathbf{H}t}$ is the unitary matrix implementing time evolution on the whole system. Now trace out $\bar{A}$:

$$\boldsymbol{\rho}_A \overset{\text{unitarily wait}}{\mapsto} \boldsymbol{\rho}_A(t) = \mathrm{tr}_{\bar{A}}\left(\mathbf{U}\boldsymbol{\rho}_A \otimes |0\rangle\langle 0|_{\bar{A}}\mathbf{U}^\dagger\right) = \sum_{i=1}^{|\bar{A}|} \langle i|\mathbf{U}|0\rangle \boldsymbol{\rho}_A \langle 0|\mathbf{U}^\dagger|i\rangle \equiv \sum_i \mathcal{K}_i \boldsymbol{\rho}_A \mathcal{K}_i^\dagger.$$

Here $\{|i\rangle\}$ is an ON basis of $\mathcal{H}_{\bar{A}}$, and we've defined *Kraus operators*

$$\mathcal{K}_i = \langle i|\mathbf{U}|0\rangle\,, \quad \sum_i \mathcal{K}_i^\dagger \mathcal{K}_i = \mathbb{1}_A\,, \quad \sum_i \mathcal{K}_i \mathcal{K}_i^\dagger = \text{whatever it wants to be.}$$

These are operators on $\mathcal{H}_A$, so this is a description of the time evolution which makes no explicit reference to $\bar{A}$ anymore. We care about the condition $\sum_i \mathcal{K}_i^\dagger \mathcal{K}_i = \mathbb{1}_A$ because it guarantees that

$$1 \overset{!}{=} \mathrm{tr}_A \boldsymbol{\rho}_A(t) = \mathrm{tr}_A \sum_i \mathcal{K}_i \boldsymbol{\rho}_A \mathcal{K}_i^\dagger \overset{\text{cyclicity of tr}}{=} \mathrm{tr}_A \underbrace{\sum_i \mathcal{K}_i^\dagger \mathcal{K}_i}_{=\mathbb{1}_A} \boldsymbol{\rho}_A = \mathrm{tr}_A \boldsymbol{\rho}_A = 1.$$

We'll see below that this parametrization is a completely general way to write a CPTP map, and the only question is to determine the Kraus operators.

Some easy examples of quantum channels:

- **Time evolution.** (unitary or subsystem),
- **Partial trace.** Time evolution takes a density matrix to another density matrix. So does ignoring part of the system. Taking partial trace is certainly trace-preserving (since you have to do the partial trace to do the whole trace). It is positive since $\mathrm{tr}_A \mathbf{S} \equiv \sum_i \langle i|_A \mathbf{S} |i\rangle_A$ is a sum of positive operators on $\bar{A}$.
- **Erasure (or reset) channel**. Quantum channels don't have to play nice:

  $$\boldsymbol{\rho} \mapsto |0\rangle\langle 0|$$

  is trace-preserving and completely positive and obliterates all information about the input state.

- **Diagonal-part channel**. Consider the channel

$$\boldsymbol{\rho} = \sum_{ij} \rho_{ij} \left|i\right\rangle \left\langle j\right| \mapsto \Phi_{QC}(\boldsymbol{\rho}) = \sum_{i} \rho_{ii} \left|i\right\rangle \left\langle i\right|$$

which keeps only the diagonal entries of the input density matrix, in some particular basis. The output is classical physics (recall that interference phenomena reside in the off-diagonal entries in the density matrix). This channel can be accomplished with $|\dim \mathcal{H}|$ Kraus operators $\mathcal{K}_i = \left|i\right\rangle \left\langle i\right|$. Notice that $\sum_i \mathcal{K}_i^\dagger \mathcal{K}_i = \mathbb{1}_{\mathcal{H}}$. And in this case $\mathcal{K} = \mathcal{K}^\dagger$, so the other order also gives $\sum_i \mathcal{K}_i \mathcal{K}_i^\dagger = \mathbb{1}$. A channel with such a set of Kraus operators is called *unital*. This condition is like the doubly-stochastic condition in the case of classical channels, and indeed also means that the uniform state $\mathbf{u} = \mathbb{1}/|\mathcal{H}|$ is a fixed point $\Phi(\mathbf{u}) = \mathbf{u}$. (In the case of Φ_{QC} above, any density matrix which is diagonal in the chosen basis is also a fixed point.)

[End of Lecture 9]

- **Phase damping channel**: A more gradual implementation of decoherence. For example, take A to be a qbit and introduce three Kraus operators

$$\mathcal{K}_0 = \sqrt{1-p}\mathbb{1}_A, \quad \mathcal{K}_1 = \sqrt{p}\left|0\right\rangle \left\langle 0\right|_A, \quad \mathcal{K}_2 = \sqrt{p}\left|1\right\rangle \left\langle 1\right|_A \ .$$

That is: with probability p it acts by the diagonal-part channel in the computational basis, and the rest of the time does nothing. So the density matrix evolves according to

$$\boldsymbol{\rho}_A \to \mathcal{E}(\boldsymbol{\rho}_A) = (1-p)\boldsymbol{\rho} + p\begin{pmatrix} \rho_{00} & 0 \\ 0 & \rho_{11} \end{pmatrix} = \begin{pmatrix} \rho_{00} & (1-p)\rho_{01} \\ (1-p)\rho_{10} & \rho_{11} \end{pmatrix}$$

Now the off-diagonal terms just shrink a little. If we do it n times

$$\boldsymbol{\rho}_A(t) = \mathcal{E}^n(\boldsymbol{\rho}_A) = \begin{pmatrix} \rho_{00} & (1-p)^n\rho_{01} \\ (1-p)^n\rho_{10} & \rho_{11} \end{pmatrix} = \begin{pmatrix} \rho_{00} & e^{-\gamma t}\rho_{01} \\ e^{-\gamma t}\rho_{10} & \rho_{11} \end{pmatrix}$$

– the off-diagonal terms decay exponentially in time $t = ndt$, like $e^{-\gamma t}$, with $\gamma = -\log(1-p)/dt \sim p/dt$, where $\mathcal{E}$ is the waiting operator for time interval dt.

Where might we obtain such Kraus operators? Suppose the environment is a 3-state system $\mathcal{H}_E = \text{span}\{\left|0\right\rangle_E, \left|1\right\rangle_E, \left|2\right\rangle_E\}$, and suppose that the result of (linear, unitary) time evolution of the coupled system over a time dt acts by

$$\begin{aligned} \mathbf{U}_{AE}\left|0\right\rangle_A \otimes \left|0\right\rangle_E &= \sqrt{1-p}\left|0\right\rangle_A \otimes \left|0\right\rangle_E + \sqrt{p}\left|0\right\rangle_A \otimes \left|1\right\rangle_E, \\ \mathbf{U}_{AE}\left|1\right\rangle_A \otimes \left|0\right\rangle_E &= \sqrt{1-p}\left|1\right\rangle_A \otimes \left|0\right\rangle_E + \sqrt{p}\left|1\right\rangle_A \otimes \left|2\right\rangle_E, \end{aligned} \tag{4.4}$$

Here's the associated poetry [from Preskill]: Suppose the two states we are considering represent positions some heavy particle in outer space, $|0\rangle_A = |x_0\rangle, |1\rangle_A = |x_1\rangle$, where x_1 and x_2 are far apart; we might like to understand why we don't encounter such a particle in a superposition $a|x_0\rangle + b|x_1\rangle$. The environment is described by *e.g.* black-body photons bouncing off of it (even in outer space, there is a nonzero background temperature associated to the cosmic microwave background). It is reasonable that these scatterings don't change the state of the heavy particle, because, lo, it is heavy. But photons scattering off the particle in different positions get scattered into different states, so the evolution of the environment should be distinct for the two different states of the heavy particle A. The probability p is determined by the scattering rate of the photons: what is the chance that a photon hits the particle in the time interval dt. Furthermore, the scattered photons go back off into space and the environment quickly resets to the state $|0\rangle_E$ with no photons and forgets about the recent little incident. This justifies the Markov approximation we made when we acted repeatedly with $\mathcal{E}$.

Completely positive trace-preserving maps. A superoperator Λ is trace-preserving (TP) if $\text{tr}_{\mathcal{H}'}\Lambda(\boldsymbol{\rho}) = \text{tr}_{\mathcal{H}}\boldsymbol{\rho}, \forall \boldsymbol{\rho}$. (It is sometimes useful to know that Λ which preserves the uniform state $\Lambda(\mathbb{1}_{\mathcal{H}}/|\mathcal{H}|) = \mathbb{1}_{\mathcal{H}'}/|\mathcal{H}'|$ is called *unital*. This second condition is like the doubly-stochastic condition on a classical channel.)

A superoperator Λ is *positive* if $\mathbf{A} \geq 0 \implies \Lambda(\mathbf{A}) \geq 0$.

$\Lambda \in \text{End}(\text{End}(\mathcal{H}_A))$ is *completely positive* (CP) if $\Lambda_A \otimes \mathbb{1}_B$ is positive $\forall \mathcal{H}_B$.

The need for complete positivity. The swap or transpose operator $T \in \text{End}(\mathcal{H}_A)$ which acts by $T(\mathbf{S}) = \mathbf{S}^\mathsf{T}$ is positive but not completely positive: Tensoring with a second copy and acting on a maximally entangled state

$$(T \otimes \mathbb{1}_B) \sum_{ij} |ii\rangle \langle jj| = \sum_{ij} |ji\rangle \langle ij|$$

produces a non-positive operator. (Notice that we had to start with an entangled state of $\mathcal{H}_{AB}$ to get a non-positive result; this is the origin of the term 'negativity' which is a measure of entanglement.)

Here's an example (really the same one) with qbits (from Schumacher appendix D): Let $\mathcal{H}_A$ be a single qbit and let $\mathcal{T}$ act on the general qbit opertor $\mathbf{A}$ by

$$\mathbf{A} = a_\mu \boldsymbol{\sigma}^\mu \equiv \sum_{\mu=0}^{3} (a_0, \vec{a})_\mu (\mathbb{1}, \vec{\boldsymbol{\sigma}})^\mu \quad \overset{\mathcal{T}}{\mapsto} \quad (a_0, a_1, a_2, -a_3)_\mu (\mathbb{1}, \vec{\boldsymbol{\sigma}})^\mu \ .$$

That is, $\mathcal{T}$ maps $\mathbb{1}, \mathbf{X}, \mathbf{Y}$ to themselves and takes $\mathcal{T}(\mathbf{Z}) = -\mathbf{Z}$. This is a positive, trace-preserving map! ($\text{tr}\mathbf{A} = 2a_0$ and $\mathbf{A} \geq 0 \Leftrightarrow a_0^2 - \vec{a}^2 \geq 0$.)

Now suppose there is another qbit B elsewhere in the world, about which our channel $\mathcal{T}$ does not care and so acts as the identity superoperator (a linear map on a tensor product is determined by its action on product states) $\mathcal{T} \otimes 1\!\!1_B$. Now consider what this map does to a Bell pair $\boldsymbol{\rho}_0 = \frac{1}{2}(|00\rangle + |11\rangle)(\langle 00| + \langle 11|)$. The definition in terms of Paulis means $\mathcal{T} : \begin{cases} |0\rangle\langle 0| \leftrightarrow |1\rangle\langle 1| \\ |0\rangle\langle 1| \leftrightarrow |0\rangle\langle 1|, \quad |1\rangle\langle 0| \leftrightarrow |1\rangle\langle 0| \end{cases}$, so the action on the maximally entangled state is

$$(\mathcal{T} \otimes 1\!\!1)(\boldsymbol{\rho}_0) = \frac{1}{2}(|10\rangle\langle 10| + |00\rangle\langle 11| + |11\rangle\langle 00| + |01\rangle\langle 01|)$$

which is of the form $\frac{1}{2}1\!\!1 \oplus \mathbf{X}$ and hence has eigenvalues $(1, 1, 1, -1)$.

The condition of complete positivity (CP) very reasonably forbids this pathology that tensoring in distant irrelevant factors in $\mathcal{H}$ can destroy positivity. And good luck finding Kraus operators that accomplish $\mathcal{T}$. (Notice that for example the very-similar-looking operation $(1\!\!1, \mathbf{X}, \mathbf{Y}, \mathbf{Z}) \to (1\!\!1, \mathbf{X}, -\mathbf{Y}, -\mathbf{Z})$ can be done with a single Kraus operator: $\boldsymbol{\rho} \to \mathbf{X}\boldsymbol{\rho}\mathbf{X}$, *i.e.* unitary evolution by $\mathbf{X}$.) We'll show later (Kraus representation theorem) that CPTP is equivalent to their existence. (If you are impatient look at Schumacher and Westmoreland's low-tech proof in appendix D.2.)

POVMs. We are used to speaking about measurements in quantum mechanics in terms of observables, namely hermitian operators $\mathcal{O} = \mathcal{O}^\dagger = \sum_a a |a\rangle\langle a|$ whose spectral representation provides a list of possible outcomes $\{a\}$ as well as a list of associated possible states in which the system ends up after measurement $\{|a\rangle\}$, which states furthermore are orthonormal and associated with orthonormal projectors

$$1\!\!1_{\mathcal{H}} = \sum_a |a\rangle\langle a| \equiv \sum_a \mathbf{P}_a; \quad \mathbf{P}_a\mathbf{P}_b = \mathbf{P}_a\delta_{ab}.$$

(The latter expressions work better than the former if there is a degeneracy in the spectrum of $\mathbf{A}$, so that the $\mathbf{P}$s are projectors of rank > 1.) The probability of obtaining outcome a when measuring $\mathbf{A}$ in state $\boldsymbol{\rho}$ is $\mathrm{tr}\boldsymbol{\rho}\mathbf{P}_a$, after which the state is $\propto \mathbf{P}_a\boldsymbol{\rho}\mathbf{P}_a$.

When our attention is focused on a subsystem of a larger system, the outcome of a measurement must be generalized somewhat. For example, suppose the whole system is in the state $\boldsymbol{\rho}_{A\bar{A}} \equiv \boldsymbol{\rho}_A \otimes |0\rangle\langle 0|_{\bar{A}}$ (where $|0\rangle_{\bar{A}}$ is some reference state of the environment $\bar{A}$) and suppose we ask for the probability to get outcome a, according to the usual rules:

$$p(a) = \mathrm{tr}_{A\bar{A}}\boldsymbol{\rho}_{A\bar{A}}\mathbf{P}_a = \mathrm{tr}\boldsymbol{\rho}\langle 0|\mathbf{P}_a|0\rangle_{\bar{A}} \equiv \mathrm{tr}\boldsymbol{\rho}M_a$$

where $M_a \equiv \langle 0|\mathbf{P}_a|0\rangle_{\bar{A}}$. In the last step we rewrote this probability without reference to the environment, in a way which has the usual form with the replacement $\mathbf{P}_a \rightsquigarrow M_a$.

The M_a are still complete:

$$\sum_a M_a = \langle 0| \sum_a \mathbf{P}_a |0\rangle_{\bar{A}} = \langle 0| \mathbb{1}_{A\bar{A}} |0\rangle_{\bar{A}} = \mathbb{1}_A$$

and they are still positive, but the price is that they are no longer orthonormal: $\mathbf{M}_a \mathbf{M}_b \neq \delta_{ab} \mathbf{M}_a$. The usual kind of measurement is called *projective measurement*, while the generalization $\{\mathbf{M}_a\}$ is called a *positive operator-valued measure* (POVM) or generalized measurement. (The particular reference state $|0\rangle_{\bar{A}}$ is not important, its purpose was merely to show us what is the form of a measurement on the subsystem.) It's not hard to show that the most general notion of measurement must take the form of a POVM. If you want some help, see Schumacher page 196.

This is a useful generalization because the lack of orthogonality of the M_a allows there to be more than $|A|$ of them.

Measurement provides another class of examples of quantum channels. If we measure the POVM $\{M_a\}$ in the state $\boldsymbol{\rho}$, and record the outcome in an extra register $R = \text{span}\{|a\rangle\}$, we can define a channel $A \to R$

$$\boldsymbol{\rho} \mapsto \sum_a \text{tr}\left(M_a \boldsymbol{\rho}\right) |a\rangle \langle a|_R .$$

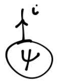
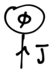
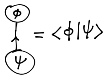
Note that a generalized measurement or POVM does not uniquely specify the state after outcome a is obtained. If we further know that upon obtaining outcome a, the state of A is $\Lambda_a(\boldsymbol{\rho})$, then we can define a channel from $A \to AR$ by

$$\boldsymbol{\rho} \mapsto \sum_a \text{tr}\left(M_a \boldsymbol{\rho}\right) \Lambda_a\left(\boldsymbol{\rho}\right) \otimes |a\rangle \langle a|_R .$$

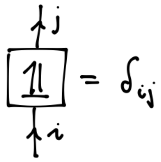
Such a channel is called an *instrument*. It can be said that we need an instrument to take a measurement.

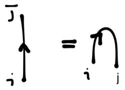
4.6 Channel duality

Feynman diagrams. The best way to understand many of the results that follow is by drawing Feynman diagrams.[25][26][27] In the context of quantum information theory and quantum computing, such diagrams are usually called quantum circuit diagrams, and a good (slightly more systematic than what I'm doing here) introduction to them can be found in the book by Schumacher. Given this translation, a better name (than Choi-Jamiolkowski Isomorphism) for what we are about to show is *channel duality.* It is exactly the same use of this term as in other fields.

- To get started, consider a state $|\psi\rangle = \sum_i \psi_i |i\rangle \in \mathcal{H}_A$. The wavefunction ψ_i is a tensor with one index which we can draw like this: . Time goes up in these diagrams – at least physicist's time in the sense that the composition of operators proceeds from bottom to top. The index is waiting to be contracted with the one on a bra vector $\langle\phi| = \sum_i \langle j| \phi_j^\star$ (which we can draw as: ) to make a number: .

- Next let's think about the object δ_{ij}, $i, j = 1...d$. We could regard this as the matrix elements of the identity operator on $\mathcal{H}_A$ of dimension $|A| = d$ (like we just used it to contract the ket and bra).



Or we could regard it as the wavefunction for (*i.e.* components in some basis of) a state in $\mathcal{H}_A \otimes \mathcal{H}_A^\star$, namely $\sum_{ij} \delta_{ij} |i\rangle \otimes |j\rangle = \sum_i |ii\rangle$. This is the statement of the isomorphism $\text{End}(\mathcal{H}_A) = \mathcal{H}_A \otimes \mathcal{H}_A^\star$. (Here the star matters if we want to respect the complex norm.) In diagrams: . Just like any Feynman diagrams, only the topology of the diagram matters. (With the one exception, also just like always, that moving an incoming line to an outgoing line costs us a complex conjugation.)

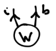
[25] In fact, for the particle physicists listening to this: the isomorphism I am about to describe is the same as the relation, shown in every dark matter talk, between direct detection and indirect detection methods.

[26] Also: perhaps you don't believe that these are the same as particle-physics Feynman diagrams because you associate the lines in those diagrams with particles, and not with tensor factors of the Hilbert space. But indeed, in a perturbatively-quantized field theory, each particle is associated with such a factor of the Hilbert space (modulo subtleties about quantum statistics) of the form

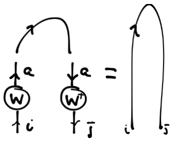
$$\mathcal{H}_{\text{particle}} \equiv \text{span}_\alpha\{|\alpha\rangle = \mathbf{c}_\alpha^\dagger |0\rangle\}$$

where α runs over whatever labels (spin, flavor, position or momentum...) the particle might carry, and $\mathbf{c}_\alpha^\dagger$ is the associated creation operator.

[27] I was not the first to notice that these diagrams are useful here. I just found this paper by Wood Biamonte and Cory which has much fancier pictures.

• Finally, let's think about a maximally entangled bipartite state on $\mathcal{H}_A \otimes \mathcal{H}_B \ni |w\rangle = \sum_{ib} w_{ib} |ib\rangle$, which looks like: 

. The statement that $|w\rangle$ is *maximally entangled* means that $\text{tr}_A |w\rangle \langle w| = \boldsymbol{\rho}_B$ and $\text{tr}_B |w\rangle \langle w| = \boldsymbol{\rho}_A$ are uniform. If $|A| = |B|$ this means they are both proportional to the identity; more generally if $|A| < |B|$, $\boldsymbol{\rho}_A = \mathbb{1}/|A|$, but $\boldsymbol{\rho}_B$ is a uniform projector onto a subspace of dimension $|A|$ inside $\mathcal{H}_B$. Let's do the same trick as above and regard w_{ia} as the coefficients of an operator $\mathbf{w} : \mathcal{H}_A^\star \to \mathcal{H}_B$. Claim: $|w\rangle$ maximally entangled $\text{tr}_B |w\rangle \langle w| = \mathbb{1}/d$ means that the operator $\mathbf{w} = w_{ia} |a\rangle \langle i|$ is an isometry $\mathbf{w}\mathbf{w}^\dagger = \mathbb{1}$, (up to the overall normalization factor) as you can easily see by diagrams at right.



[I found the discussion by Wolf to be very useful for the following, which is a warm-up for the channel duality theorem.]

• Here is a fun and mind-bending application of the maximally entangled state $|\Phi\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} |ii\rangle$. Let me call the second factor of the Hilbert space $\mathcal{H}_C$ and assume $|C| \geq |A| \equiv d$. It can be called **Schrödinger lemma**: Consider a bipartite state $|\psi\rangle \in \mathcal{H}_{AC}$ with $\boldsymbol{\rho}_C = \text{tr}_A |\psi\rangle \langle \psi|$. *Any* such state can be made from $|\Phi\rangle$ without doing anything to A:

$$|\psi\rangle = (\mathbb{1} \otimes R) |\Phi\rangle \qquad R = \sqrt{d \boldsymbol{\rho}_C} V$$

where V is an isometry. The key point is that for any unitary on A,

$$|\Phi\rangle = \mathbf{U} \otimes \left(\mathbf{U}^\star \oplus \mathbb{1}_{|C|-|A|} \right)_C |\Phi\rangle .$$

Again this is easiest in terms of diagrams.

Finite condition for CP. A reason to care about the preceding result is that it can be used to find a criterion for complete positivity: $\mathcal{E} : \text{End}(A) \to \text{End}(D)$ is CP IFF

$$(\mathcal{E} \otimes \mathbb{1}_d) (|\Phi\rangle \langle \Phi|) \geq 0 \tag{4.5}$$

where the spectator factor has the same dimension as A.

Proof: $\boxed{\Longleftarrow}$ follows from the the definition of CP. To see $\boxed{\Longrightarrow}$, take any state $\boldsymbol{\rho} \in \text{End}(A \otimes B)$ on which we might hope $\mathcal{E} \otimes \mathbb{1}_B$ is positive. This desideratum $(\mathcal{E} \otimes \mathbb{1}_B) (\boldsymbol{\rho} = \sum_k p_k |k\rangle \langle k|) \geq 0$ will follow if it's true for every 1d projector $|k\rangle \langle k|$ in the spectral representation of $\boldsymbol{\rho}$:

$$0 \leq (\mathcal{E} \otimes \mathbb{1}_B) (|k\rangle \langle k|) . \tag{4.6}$$

But now the Schrödinger lemma says we can write

$$|k\rangle = \mathbb{1}_d \otimes R_k |\Phi\rangle$$

for some map $R_k \in \text{Hom}(C, B)$, where C is the auxiliary space from the discussion above. But then

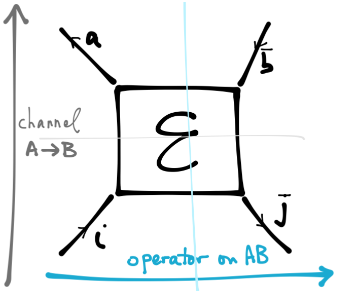
$$\begin{aligned}(\mathcal{E} \otimes 1\!\!1_B)\left(|k\rangle\langle k|\right) &= (\mathcal{E} \otimes 1\!\!1_B)\left(1\!\!1_d \otimes R_k |\Phi\rangle\langle\Phi| 1\!\!1_d \otimes R_k^\dagger\right) \\ &= 1\!\!1_d \otimes R_k \left[(\mathcal{E} \otimes 1\!\!1_B)\left(|\Phi\rangle\langle\Phi|\right)\right] 1\!\!1_d \otimes R_k^\dagger \geq 0 \end{aligned} \tag{4.7}$$

where the penultimate step used the placement of the identity operators, and the last step follows from our hypothesis (4.5) since $B \to ABA^\dagger$ preserves positivity[28], and we have (4.6). ■ [End of Lecture 10]

C-Jam Lemma: (Choi-Jamiolkowski isomorphism) [Christaindl, lecture 5, Renner §4.4.2] The summary is: the set of quantum channels $A \to B$ is the same as the set of maximally entangled states of AB. To make this more precise, consider a superoperator

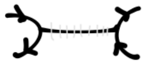
$$\begin{aligned}\mathcal{E} : \text{End}(A) &\to \text{End}(B) \\ \boldsymbol{\rho}_A &\mapsto \mathcal{E}(\boldsymbol{\rho}_A)\ . \\ \boldsymbol{\rho}_{ij} &\mapsto \mathcal{E}_{ab}^{ij}\boldsymbol{\rho}_{ij}\end{aligned}$$

ij are indices on (*i.e.* labels on an ON basis of) $\mathcal{H}_A$ and ab are indices on $\mathcal{H}_B$. In thinking of $\mathcal{E}$ as a channel $A \to B$, we regard the 4-index object $\mathcal{E}_{ab}^{ij}$ as a matrix with multi-indices ab and ij. Now just look at it sideways (as in the figure at right). That is, regard the 4-index object $\mathcal{E}_{ab}^{ij}$ as a matrix with multi-indices ai and bj. Lo, it is now an element of End(AB), an operator on AB.

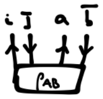


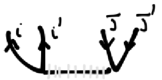
A quantum channel is not just any linear map, and a density matrix is not just any operator. We can make the channel-duality statement more precise by introducing a second Hilbert space isomorphic to $\mathcal{H}_A \simeq \mathcal{H}_{A'}$. Such an isomorphism $\delta : \mathcal{H}_A \to \mathcal{H}_{A'}$ specifies a *maximally entangled state* of A and A'

$$|\delta\rangle \equiv \sum_{ij} \delta_{ij} |i\rangle_A \otimes |j\rangle_{A'} = \sum_i |ii\rangle\ .$$

(Note that I didn't normalize it.) Maximally entangled means $\text{tr}_{A'} |\delta\rangle\langle\delta| \propto 1\!\!1_A$. The density matrix for the maximally entangled state looks like this (time goes up): $|\delta\rangle\langle\delta| =$  . (The dashed line is meaningless and could be omitted.) We are going to freely use the isomorphisms described above now, so a density matrix on AB looks

[28] B is positive means that $\langle v|B|v\rangle \geq 0 \forall |v\rangle$. But then $\langle v|ABA^\dagger|v\rangle \overset{|\tilde{v}\rangle \equiv A|v\rangle}{=} \langle\tilde{v}| B |\tilde{v}\rangle \geq 0$ as well.

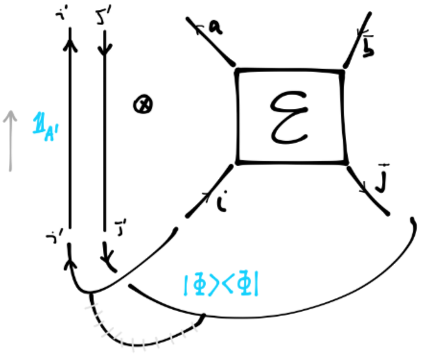
like this:  In particular, the density matrix for the pure state $|\delta\rangle$ can also be

drawn like  .

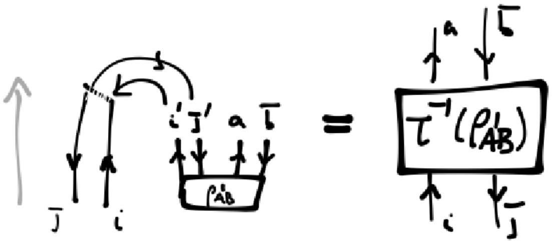
Then we can state the C-JAM result in terms of the linear map

$$\tau : \begin{array}{ccc} \mathrm{Hom}(\mathrm{End}(A),\mathrm{End}(B)) \to & & \mathrm{End}(A'B) \\ \mathcal{E} & & \mapsto \tau(\mathcal{E}) = (\mathcal{E}\otimes 1\!\!1_{A'})\left(\frac{|\delta\rangle\langle\delta|}{d}\right) \end{array}$$

That this is a vector space isomorphism we can prove by the following diagram (which should be read from bottom to top):



Its inverse is the following: given an operator $\boldsymbol{\rho}_{A'B}$ on $A'B$, make a channel $A \to B$ using only $\boldsymbol{\rho}$ and $|\delta\rangle$. There is only one way to attach the indices:



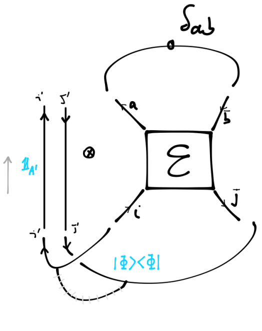
In equations it's (a bit trickier for me):

$$\tau^{-1} : \boldsymbol{\rho}_{A'B} \mapsto \left(\mathbf{X}_A \mapsto d\mathrm{tr}_{A'}\left(\mathbf{X}_{A'}^T \otimes 1\!\!1_B \boldsymbol{\rho}_{A'B}\right)\right)$$

The thing on the RHS of the map is a map from operators on A to operators on B. Here we used the isomorphism δ between A and A'.

It is easiest to check with the diagrams that indeed: $\tau \circ \tau^{-1} = 1\!\!1_{\mathrm{End}(A'B)}$ (and the other order, too). The more nontrivial bit is the claim that τ maps quantum channels $\mathsf{CPTP}(A \to B)$ to density matrices on $A'B$ (and specifically, it is an isomorphism with maximally entangled density matrices on $A'B$). $\mathcal{E}$ is CP guarantees that the density matrix $\tau(\mathcal{E})$ is positive by the definition of CP. And $\mathcal{E}$ is trace-preserving means $1 = \mathrm{tr}_B\mathcal{E}(\boldsymbol{\rho}_A) = \sum_a \mathcal{E}^{ij}_{aa}\rho_{ij}\forall\rho_{ij}|\sum_i \rho_{ii} = 1$. But in particular it is true for $\rho = 1\!\!1/d$ which is what we need for $1 = \mathrm{tr}_{A'B}\tau(\mathcal{E}) = \sum_{ai}\mathcal{E}^{ii}_{aa}$.

Now, any density matrix in the image of τ is maximally entangled. This is because it has $\mathrm{tr}_B \boldsymbol{\rho}_{A'B} = \mathbb{1}_{A'}/d$, as you can see from the diagrams by contracting the a and $\bar{b}$ indices – this gives $\mathrm{tr}_B \tau(\mathcal{E})^{i'j'} = \mathcal{E}^{i'j}_{aa}$ which must be $d\delta^{i'j'}$ since $\mathcal{E}$ is trace-preserving (*i.e.* $\mathcal{E}^{ij}_{aa}\rho_{ij} = 1$ for any normalized ρ).

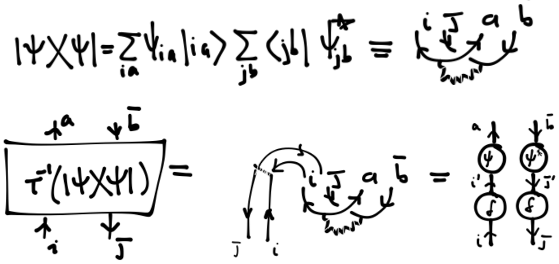


And every such density matrix on $A'B$ is in the image of τ. The image of unitary evolution (actually $\mathbf{U}$ is an isometry if $|A| \neq |B|$) is a pure state:

$$\tau(\boldsymbol{\rho} \to \mathbf{U}\boldsymbol{\rho}\mathbf{U}^\dagger) = \mathbf{U} \otimes \mathbb{1} \left|\delta\right\rangle \left\langle\delta\right| \mathbf{U}^\dagger \otimes \mathbb{1}$$

(For example, the image of the identity channel is the state $\left|\delta\right\rangle \left\langle\delta\right|/d$.)

Conversely, the pre-image of any pure state $\left|\psi\right\rangle = \psi_{ia}\left|ia\right\rangle$ (which must be maximally mixed on A' to have a pre-image – this is why ψ_{ia} is an isometry) is such an isometric evolution. The general pre-image is then a convex combination of conjugation by isometries which is completely positive (since it is a Kraus representation).


$|\psi\rangle\langle\psi| = \sum_{ia} \psi_{ia} |ia\rangle \sum_{jb} \langle jb| \psi^*_{jb} =$
i
$\bar{j}$
a
$\bar{b}$
$\tau^{-1}(|\psi\rangle\langle\psi|) =$
a
$\bar{b}$
i
$\bar{j}$
$=$
ψ
ψ^*
δ
i'
$\bar{j}'$

• Moving outside the set of CP maps, the condition that the operator $\tau(\mathcal{E})$ is hermitian is that $\mathcal{E}$ is hermiticity-preserving $\mathcal{E}(\mathbf{A}^\dagger) = \mathcal{E}(\mathbf{A})^\dagger$.

• The condition that $\mathcal{E}$ is unital $\mathcal{E}(\mathbb{1}) = \mathbb{1}$ is that $\mathrm{tr}_{A'}\tau(\mathcal{E}) = \frac{1}{|B|}\mathbb{1}_B$ is the identity on B.

Application of C-Jam Isomorphism: Let $\mathcal{M}$ be an *instrument*, as we defined earlier. With a little repackaging, this is a set of CP maps $\mathcal{M}_\alpha$ whose sum is trace-preserving $\mathrm{tr}\sum_\alpha \mathcal{M}_\alpha(\boldsymbol{\rho}) = \mathrm{tr}\boldsymbol{\rho}$. The label α is the measurement outcome, which obtains with probability $p(\alpha) = \mathrm{tr}\mathcal{M}_\alpha(\boldsymbol{\rho})$. It is tempting to say that when the outcome is α, the resulting state is $\mathcal{M}_\alpha(\boldsymbol{\rho})/p(\alpha)$.

No information without disturbance: if on average, there is no disturbance of the state, $\sum_\alpha \mathcal{M}_\alpha = \mathbb{1}$, then $\mathcal{M}_\alpha \propto \mathbb{1}\ \forall\alpha$ (and $p(\alpha)$ is independent of $\boldsymbol{\rho}$).

Proof: the image under C-Jam of the BHS of the equation $\mathbb{1} = \sum_\alpha \mathcal{M}_\alpha$ is $\left|\Phi\right\rangle\left\langle\Phi\right| = \sum_\alpha \tau(\mathcal{M}_\alpha)$. Since each $\tau(\mathcal{M}_\alpha) \geq 0$, this is a convex decomposition of a pure state. We'll prove in a moment ((4.11)) that this means every term is itself proportional to the pure state: $\tau(\mathcal{M}_\alpha) = c_\alpha \left|\Phi\right\rangle\left\langle\Phi\right|, c_\alpha \geq 0$. The inverse of C-Jam then says $\mathcal{M}_\alpha = c_\alpha\mathbb{1}$, and $p(\alpha) = c_\alpha$ for any state $\boldsymbol{\rho}$.

4.7 Purification, part 2

The notion of purification is the hero of this subsection.

4.7.1 Concavity of the entropy

[C&N p. 517] A convex combination of density matrices is a density matrix:

$$\sum_i p_i \boldsymbol{\rho}_i \equiv \boldsymbol{\rho}_{\text{av}} ~,$$

where $\{p_i\}$ is a probability distribution on i ($p_i \geq 0, \sum_i p_i = 1$). How does the vN entropy behave under such averages? It is concave:

$$S\left(\boldsymbol{\rho}_{\text{av}}\right) \geq \sum_i p_i S(\boldsymbol{\rho}_i) ~. \tag{4.8}$$

This statement seems reasonable since on the LHS we have the extra uncertainty about the value of the label i.

Proof of (4.8): The proof uses a (partial) purification. Suppose each $\boldsymbol{\rho}_i \in \text{End}(A)$. Introduce an auxiliary system B with $\mathcal{H}_B \supset \text{span}\{|i\rangle\}_{\text{ON}}$ which we will use to store the value of the label i. Take

$$\boldsymbol{\rho}_{AB} \equiv \sum_i p_i \boldsymbol{\rho}_i \otimes |i\rangle \langle i| \,. \tag{4.9}$$

Simple calculations give $\boldsymbol{\rho}_A \equiv \text{tr}_B \boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_{\text{av}}$ and hence

$$S(\boldsymbol{\rho}_A) = S\left(\boldsymbol{\rho}_{\text{av}}\right), \quad \textit{and} \;\; S\left(\boldsymbol{\rho}_B\right) = S\left(\sum_i p_i \left|i\right\rangle \left\langle i\right|\right) = H(p).$$

Introduce a spectral decomposition of each $\boldsymbol{\rho}_i = \sum_j \lambda_j^{(i)} \left|e_j^{(i)}\right\rangle \left\langle e_j^{(i)}\right|$. (These eigenvectors are ON (and $\sum_j \lambda_j^{(i)} = 1$) *for each* i but since the $\boldsymbol{\rho}_i$ need not commute are different bases for each i! Beware!) So $\boldsymbol{\rho}_{\text{av}} = \sum_i \sum_j p_i \lambda_j^{(i)} \left|e_j^{(i)}\right\rangle \left\langle e_j^{(i)}\right|$ but this is not the spectral representation of $\boldsymbol{\rho}_{\text{av}}$. However,

$$\boldsymbol{\rho}_{AB} = \sum_i \sum_j p_i \lambda_j^{(i)} \left|e_j^{(i)}\right\rangle \left\langle e_j^{(i)}\right| \otimes |i\rangle\langle i|$$

is the spectral representation of $\boldsymbol{\rho}_{AB}$, because the states $\left|e_j^{(i)}\right\rangle \otimes |i\rangle$ *are* orthonormal for all i, j. Then

$$S(\boldsymbol{\rho}_{AB}) = -\sum_i p_i \sum_j \lambda^{(i)_j} \log\left(p_i \lambda_j^{(i)}\right) = H(p) + \sum_i p_i S(\boldsymbol{\rho}_i).$$

Subadditivity of the vN entropy is

$$\begin{aligned} S(AB) &\leq S(A) + S(B) \\ \sum_i p_i S(\boldsymbol{\rho}_i) + H(p) &\leq S(\boldsymbol{\rho}_{\mathrm{av}}) + H(p) \end{aligned} \tag{4.10}$$

which is the concavity condition. $\boxed{4.8}$

The subadditivity inequality is saturated IFF $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$ (since $S(A)+S(B)-S(AB) = I(A:B) = D(\boldsymbol{\rho}_{AB}||\boldsymbol{\rho}_A\boldsymbol{\rho}_B)$ which vanishes only when the two states are the same). This only happens if the $\boldsymbol{\rho}_i$ are all equal to $\boldsymbol{\rho}_{\mathrm{av}}$.

Concavity of the entropy is equivalent to the statement that the *Holevo quantity*

$$\chi(p_i, \boldsymbol{\sigma}_i) \equiv S(\boldsymbol{\sigma}_{\mathrm{av}}) - \sum_i p_i S(\boldsymbol{\sigma}_i)$$

is positive $\chi \geq 0$. This quantity is very useful in the study of transmission of classical information with quantum channels, more below.

Pure states are extremal. Here is a simple application of the concavity of the entropy: In a convex decomposition of a pure state, every term is proportional to the state itself:

$$|\Phi\rangle\langle\Phi| = \sum_a p_a \boldsymbol{\rho}_a \leftrightarrow \boldsymbol{\rho}_a \propto |\Phi\rangle\langle\Phi| \ . \tag{4.11}$$

Here's the proof:

$$0 = S\left(|\Phi\rangle\langle\Phi|\right) = S\left(\sum_a p_a \boldsymbol{\rho}_a \leftrightarrow \boldsymbol{\rho}_a\right) \overset{(4.8)}{\leq} \sum_a p_a S(\boldsymbol{\rho}_a) \geq 0$$

and therefore each term on the RHS must vanish. Once we know each $\boldsymbol{\rho}_a$ is pure, they are all proportional to $|\Phi\rangle\langle\Phi|$ because the p_a are positive.

We conclude that the pure states lie at the boundaries of the (convex!) set of density matrices.

By the way, there is a classical analog of this statement: consider a collection of probability distributions π^α on a random variable X, so $\sum_x \pi^\alpha_x = 1, \pi^\alpha_x \geq 0, \ \forall x$. Then a convex combination of these $\pi_{\mathrm{av}} \equiv \sum_\alpha p_\alpha \pi^\alpha$ is also a probability distribution on X. And indeed the entropy of the average distribution is larger than the average of the entropies:

$$H(\pi_{\mathrm{av}}) \geq \sum_\alpha p_\alpha H(\pi^\alpha)$$

as you'll check on the homework.

Concavity is a *lower* bound on $S(\boldsymbol{\sigma}_{\text{av}})$. There is also an upper bound [C&N Theorem 11.10]:

$$S(\boldsymbol{\sigma}_{\text{av}}) \leq \sum_i p_i S(\boldsymbol{\sigma}_i) + H(p). \tag{4.12}$$

In the classical analog, this inequality is always saturated.

Proof of (4.12): Here is the proof, first for the case where the $\boldsymbol{\sigma}_i$ are pure states, $\boldsymbol{\sigma}_i = |\psi_i\rangle\langle\psi_i|$. Define a purification (surprise, surprise) of $\boldsymbol{\sigma}_{\text{av}}$, $|AB\rangle = \sqrt{p_i}\,|\psi_i\rangle \otimes |i\rangle_B$ where the $|i\rangle_B$ are ON (even though the $|\psi_i\rangle$ need not be). Purity of the whole system says $S(B) = S(A) = S(\boldsymbol{\sigma}_{\text{av}})$. But now let's consider measuring the observable $|i\rangle\langle i|$ on B; the resulting probability distribution on i is just p_i. We proved (in (4.2)) that the Shannon entropy of the distribution resulting from a measurement is bigger than the initial vN entropy[29] this result shows that the entropy :

$$H(p) \geq S(B) = S(\boldsymbol{\sigma}_{\text{av}})$$

which is (4.12) for this special case (since $S(|\psi_i\rangle\langle\psi_i|) = 0, \forall i$).

To do the general case, make again a spectral decomposition of each $\boldsymbol{\sigma}_i = \sum_j \lambda_j^{(i)} \left|e_j^{(i)}\right\rangle \left\langle e_j^{(i)}\right|$. Although $\boldsymbol{\sigma}_{\text{av}} = \sum_i \sum_j p_i \lambda_j^{(i)} \left|e_j^{(i)}\right\rangle \left\langle e_j^{(i)}\right|$ is not the spectral representation of $\boldsymbol{\sigma}_{\text{av}}$, the numbers $\{p_i \lambda_j^{(i)}\}$ do provide a probability distribution on the set $\{ij\}$. So we can just apply the pure-state result above with $p_i \rightsquigarrow p_i \lambda_j^{(i)}$ and $|\psi_i\rangle \rightsquigarrow \left|e_j^{(i)}\right\rangle$. So we have

$$\begin{aligned} S(\boldsymbol{\sigma}_{\text{av}}) \leq H\left(p_i \lambda_j^{(i)}\right) &= -\sum_{ij} p_i \lambda_j^{(i)} \log\left(p_i \lambda_j^{(i)}\right) \\ &= -\sum_i p_i \log p_i - \sum_i p_i \sum_j \lambda_j^{(i)} \log \lambda_j^{(i)} = H(p) + \sum_i p_i S(\boldsymbol{\sigma}_i). \end{aligned}$$

The upper bound is saturated IFF the $\boldsymbol{\sigma}_i$ have orthogonal support. $\boxed{4.12}$

Summary:

$$0 \leq \chi(p_i, \boldsymbol{\rho}_i) \leq H(p)$$

– the left inequality is saturated if $\boldsymbol{\rho}_i = \boldsymbol{\rho}_{\text{av}} \forall i$, and the right is saturated if $\boldsymbol{\rho}_i \perp \boldsymbol{\rho}_j$.

[29] Actually, since the state of B after such a projective measurement of $1\!\!1_A \otimes |i\rangle\langle i|$ is $\boldsymbol{\rho}'_B = \sum_i p_i |i\rangle\langle i|$, whose vN entropy is $S(\boldsymbol{\rho}'_B) = H(p)$, we see that projective measurement increases the entropy (if we don't look at the outcome).

4.7.2 Stinespring dilation and Kraus representation.

Every CPTP map can be regarded as a unitary on some larger Hilbert space (followed by partial trace). This larger evolution is called a *dilation.*

Low-tech dilatation. If we are *given* Kraus operators for our channel $\{K_i\}$, the dilatation is easy: define the map

$$|\psi\rangle \otimes |0\rangle_E \mapsto \sum_i K_i |\psi\rangle \otimes |i\rangle_E$$

where $|i\rangle_E$ is an ON basis of some ancillary space. Then we can find a unitary which acts this way on this particular subspace. And the Kraus operators are related to it as above, $K_i = \langle i| K_i |0\rangle_E$. [End of Lecture 11]

To see that this is the case in general, first we show: Any quantum channel $\mathcal{E}$: $\mathrm{End}(A) \to \mathrm{End}(B)$ can be written as

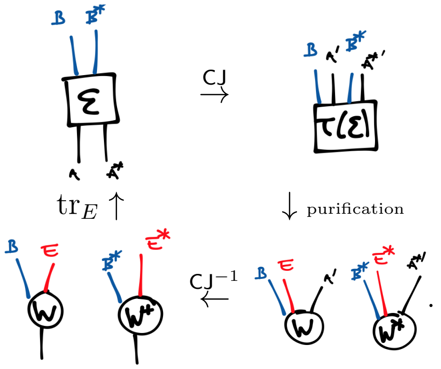
$$\mathbf{X} \mapsto \mathcal{E}(\mathbf{X}) = \mathrm{tr}_E \mathbf{W}\left(\mathbf{X}\right)\mathbf{W}^\dagger$$

for isometries $\mathbf{W} \in \mathrm{Hom}(A, BE)$.

Proof: the following diagram commutes:

$$\begin{tikzcd}
\mathcal{E}_{A\to B} \arrow[r, "\mathsf{CJ}"] & \boldsymbol{\rho}_{A'B} \arrow[d, "\text{purification}"] \\
\mathcal{W}_{A\to BE} \arrow[u, "\mathrm{tr}_E"] & |W\rangle\langle W|_{A'BE} \arrow[l, "\mathsf{CJ}^{-1}"']
\end{tikzcd}$$

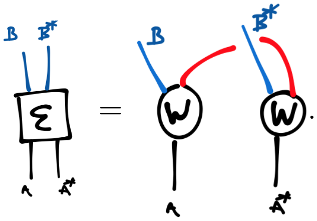
In pictures, this is:



The channel $\mathcal{W}_{A\to BE}$ acts by

$$\boldsymbol{\rho}_A \to \mathcal{W}_{A\to BE}(\boldsymbol{\rho}_A) = \mathbf{W}\boldsymbol{\rho}_A\mathbf{W}^\dagger$$

where $\mathbf{W} : A \to BE$, $\mathbf{W}^\dagger\mathbf{W} = \mathbb{1}_A$ is the isometry made by CJAM^{-1} from the pure state $|W\rangle_{A'BE}$. The final result is



For the special case of channels acting on a fixed system (*i.e.*A to A) we can turn this into unitary evolution: Any quantum channel $\mathcal{E} : \text{End}(A) \to \text{End}(A)$ can be written as

$$\mathbf{X} \mapsto \mathcal{E}(\mathbf{X}) = \text{tr}_E \mathbf{U} \left(\mathbf{X} \otimes |0\rangle \langle 0|_E\right) \mathbf{U}^\dagger$$

for unitaries $\mathbf{U}$ on AE. This we do just by filling in the missing entries of $\mathbf{U}$, just as we did in the easy dilation result.

Kraus representation theorem. This follows immediately by picking a basis $\{|i\rangle, i = 1..r\}$ for E in the previous result:

$$\mathcal{E}(\mathbf{X}) = \text{tr}_E \mathbf{W}\mathbf{X}\mathbf{W}^\dagger = \sum_{i=1}^{r} \langle i| \mathbf{W}\mathbf{X}\mathbf{W}^\dagger |i\rangle = \sum_i \mathcal{K}_i \mathbf{X} \mathcal{K}_i^\dagger$$

with

$$\mathcal{K}_i \equiv \langle i|_E \mathbf{W}_{A\to BE} : A \to B.$$

Notice that there is no need to choose a reference state of the environment.

Some comments about Kraus (or operator-sum) representations of channels which I could have made earlier but which will be clearer now:

$\mathcal{E}$ is TP $\leftrightarrow \sum_i \mathcal{K}_i^\dagger \mathcal{K}_i = \mathbb{1}$. $\mathcal{E}$ is unital $\leftrightarrow \sum_i \mathcal{K}_i \mathcal{K}_i^\dagger = \mathbb{1}$.

For any channel $\mathcal{E} : \text{End}(A) \to \text{End}(B)$ we can define the *adjoint channel* $\mathcal{E}^\ddagger$: $\text{End}(B) \to \text{End}(A)$ by

$$\text{tr}_B \left(\mathbf{B}\mathcal{E}(\mathbf{A})\right) = \text{tr}_A \left(\mathcal{E}^\ddagger(\mathbf{B})\mathbf{A}\right)$$

for any two Hermitian operators on A and B. Note that the adjoint here gets a weird dagger, since it is adjoint (on superoperators!) with respect to the Hilbert-Schmidt inner product on operators, not the ordinary Dirac inner product on vectors. Happily, though, the Kraus operators of the adjoint channel are $\{\mathcal{K}_i^\dagger\}$:

$$\text{tr}_B \boldsymbol{\rho}_B \mathcal{E}(\boldsymbol{\rho}_A) = \text{tr}_B \boldsymbol{\rho}_B \sum_i \mathcal{K}_i \boldsymbol{\rho}_A \mathcal{K}_i^\dagger = \sum_i \text{tr}_A \mathcal{K}_i^\dagger \boldsymbol{\rho}_B \mathcal{K}_i \boldsymbol{\rho}_A = \text{tr}_A \mathcal{E}^\ddagger(\boldsymbol{\rho}_B)\boldsymbol{\rho}_A$$

where the middle step uses cyclicity of the trace.

This is a different notion of channel duality from the C-Jam duality! The previous two conditions (TP and unital) are 'dual' (actually adjoint) in this sense, *e.g.* $\mathcal{E}^\ddagger(\mathbb{1}) = \mathbb{1}$ means $\mathcal{E}$ is TP and vice versa.

The number r of Kraus operators is called the *Kraus rank* of $\mathcal{E}$. It is the Schmidt rank of $\mathsf{CJ}(\mathcal{E})$. Note that it is *not* the rank of $\mathcal{E}$ as a linear map. For example, $\mathcal{E} = \mathbb{1}$ has full rank, but Kraus rank $r = 1$, while the trace map $\mathcal{E}(B) = \text{tr}(B)$ has rank 1 (the image is one-dimensional) but Kraus rank d.

The representation is not unique, since we can rotate the environment: $\{\mathcal{K}\} \simeq \{\tilde{\mathcal{K}}\}$ produce the same channel iff $\mathcal{K}_k = \sum_l u_{kl}\tilde{\mathcal{K}}_l$ where u_{kl} is a unitary matrix in the kl indices.

It is possible to choose a non-minimal Kraus representation with extra Kraus operators. It is, however, possible (by Gram-Schmidt on the environment) to choose $\text{tr}\mathcal{K}_i^\dagger \mathcal{K}_j \propto \delta_{ij}$.

4.8 Deep facts

So far the entropy bounds we've discussed have not involved any heavy lifting. Now we come to the hard stuff, associated with *strong subadditivity* (SSA). It is quite remarkable how many interesting statements can be shown to be equivalent to SSA by relatively simple operations; to get to any of them requires a step which seems relatively difficult. It is like a mountain plateau. Or maybe like a particular circle of hell. (This point is subjective in many ways, but I suspect there is some objective truth in there.)

The most memorable-to-me of these statements is:

(1) **Monotonicity of Relative entropy** (under trace): Given two states $\boldsymbol{\rho}_{AB}, \boldsymbol{\sigma}_{AB}$ on $\mathcal{H}_A \otimes \mathcal{H}_B$,

$$\boxed{D(\boldsymbol{\rho}_{AB}||\boldsymbol{\sigma}_{AB}) \geq D(\boldsymbol{\rho}_A||\boldsymbol{\sigma}_A)\ .} \tag{4.13}$$

In words: forgetting about B can only decrease the distance between these states. Before proving this, let's derive some corollaries (there are like a million equivalent statements):

(2) **Strong subadditivity** (of the vN entropy): Consider a tripartite system ABC and let $\boldsymbol{\rho} = \boldsymbol{\rho}_{ABC}$ while $\boldsymbol{\sigma} = \boldsymbol{\rho}_B \otimes \boldsymbol{\rho}_{AC}$ is the product of the marginals. Then using (4.13), forgetting C, says that discarding a part of the system cannot increase the mutual information:

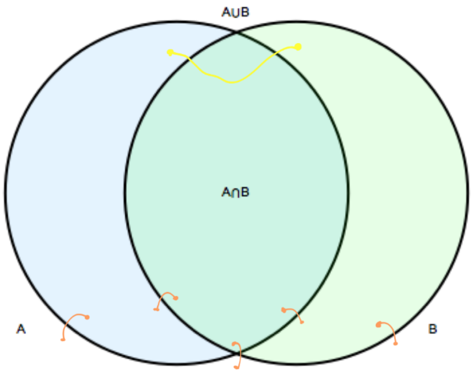
$$\begin{aligned} S(B:AC) &\geq S(B:A). \\ S(B) + S(AC) - S(ABC) &\geq S(A) + S(B) - S(AB) \\ S(AC) + S(AB) &\geq S(A) + S(ABC) \end{aligned} \tag{4.14}$$

The last of these (identical) statements is called *strong sub-additivity* (SSA). (It is *strong* at least in the sense that it implies subadditivity by taking $A =$ nothing.)

A relabeling translates SSA to a statement about inclusion and exclusion:

$$S(A \cup B) + S(A \cap B) \leq S(A) + S(B). \tag{4.15}$$

At right is a heuristic (mnemonic?) I learned from Tarun Grover. For definiteness consider the case where A, B are the Hilbert spaces associated with subregions of space occupied by an extensive quantum system. Suppose the whole system is in pure state, so that the entropy of the reduced states of $A, B, A \cup B, A \cap B$ all arise from entanglement with their respective complements. The heuristic arises by visualize this entanglement as singlet bonds, in the same way that we can denote a maximally entangled state of two qbits $|\uparrow_1\downarrow_2\rangle - |\downarrow_1\uparrow_2\rangle$ by joining them with a line $(1-2)$. Now, if we draw a singlet between each region and each of the parts of its complement, and count singlets, we see that most of them (the orange ones) contribute to the BHS of (4.15), but the bond between $A \setminus B$ and $B \setminus A$ (in yellow) contributes only to the RHS (actually twice).



Another version of SSA is

$$S(A) + S(B) \leq S(AC) + S(BC). \tag{4.16}$$

This can be proved using (4.14) by yet another purification move (see the homework).

Recall that the (not so hard) proof of the classical version of this statement (for Shannon entropies) which you found on the homework relied on the existence of (positive) conditional entropies like $H(B|C)$. (What is the quantum version of a conditional probability? It's a channel.) We can still define $S(B|C) \equiv S(BC) - S(C)$, but it is negative if $S(BC)$ is more entangled between B and C than with its environment, *i.e.* when the state is very quantum. Nevertheless, it is still common to call the deviation from saturation of SSA the conditional mutual information:

$$I(A:C|B) \equiv S(A:CB) - S(A:B) \geq 0 \ .$$

When this condition is saturated, ABC are said to form a quantum Markov chain. Roughly, it means that C and A only talk to each other through B. More later on this.

If we are willing to call $S(A|B) \equiv S(AB) - S(B)$ despite the fact that it can be negative, then another statement of SSA is:
conditioning decreases entropy: $S(A|BC) \leq S(A|B)$.

(3) Finally, one more statement which is nontrivially equivalent to SSA is the concavity of the conditional entropy $S(A|B)$ as a function of $\boldsymbol{\rho}_{AB}$.

$\boxed{\Longleftarrow}$ This statement implies SSA in the following clever way (C&N p.521): It implies that the function

$$T(\boldsymbol{\rho}_{ABC}) \equiv -S(C|A) - S(C|B)$$

is a convex function of $\boldsymbol{\rho}_{ABC} = \sum_i p_i \left|i\right\rangle \left\langle i\right|$. Now feed this spectral representation (which for a density matrix is a convex decomposition) into T:

$$T(\boldsymbol{\rho}_{ABC}) \overset{\text{convex}}{\leq} \sum_i p_i T(\left|i\right\rangle \left\langle i\right|).$$

But for a pure state on ABC, $S(AC) = S(B)$ and $S(BC) = S(A)$, so $T(\text{pure}) = 0$. Therefore

$$0 \geq T(\boldsymbol{\rho}_{ABC}) = S(A) + S(B) - S(AC) - S(BC)$$

which is a version of SSA in the form (4.16).

$\boxed{\Longrightarrow}$ To see that SSA implies concavity of the conditional entropy: Since

$$D(\boldsymbol{\rho}_{AB}||\mathbb{1}/d \otimes \boldsymbol{\rho}_B) = -S(AB) + S(B) + \log d = -S(A|B) + \log d$$

concavity of $S(A|B)$ follows from SSA with one extra trick which you'll get to use on HW 7.

I'll give a bit more of a guide to the byroads winding through this particular circle of hell below; if you are impatient, see §5.3 of this very clear paper of Ruskai.

Before proving any of these statements, let me try to convince you that it is worthwhile. In particular, let's consider consequences of combining them with the purification idea. The Stinespring dilation theorem tells us that any channel is purification, unitary evolution, partial trace. But the entropies we are discussing are basis-independent, and hence do not change upon unitary evolution of the whole space. This has the immediate consequence that the relative entropy is monotonic not just under partial trace, but under *any channel*:

$$D(\boldsymbol{\rho}||\boldsymbol{\sigma}) \geq D(\mathcal{E}(\boldsymbol{\rho})||\mathcal{E}(\boldsymbol{\sigma})). \tag{4.17}$$

More explicitly: the effects of the channel on our system S can be reproduced by introducing an ancillary environment E, initially in some reference product state with S, $\mathbf{P}_E = \left|0\right\rangle \otimes \left\langle 0\right|_E$; then unitarily evolving the whole system SE, then throwing away E. The operation of appending E does not change the relative entropy:

$$D(\boldsymbol{\rho}||\boldsymbol{\sigma}) = D(\boldsymbol{\rho} \otimes \mathbf{P}_E||\boldsymbol{\sigma} \otimes \mathbf{P}_E).$$

Neither does unitary evolution on SE :

$$D(\mathbf{U}\boldsymbol{\rho}_{SE}\mathbf{U}^\dagger||\mathbf{U}\boldsymbol{\sigma}_{SE}\mathbf{U}^\dagger) = D(\boldsymbol{\rho}_{SE}||\boldsymbol{\sigma}_{SE}).$$

The only step that does anything is tracing out E, which is governed by our previous monotonocity result, equivalent to SSA. $\boxed{4.17}$

In particular, a quantum channel cannot increase the mutual information

$$I_{\boldsymbol{\rho}}(A:B) = D(\boldsymbol{\rho}_{AB}||\boldsymbol{\rho}_A\boldsymbol{\rho}_B) \geq D(\mathcal{E}(\boldsymbol{\rho}_{AB})||\mathcal{E}(\boldsymbol{\rho}_A\boldsymbol{\rho}_B)) = I_{\mathcal{E}(\boldsymbol{\rho})}(A:B).$$

So these can be called quantum data processing inequalities.

Holevo bound. Another application of the above deep facts is a bound on the information-transmitting capacity of protocols like quantum teleportation and dense coding. More specifically, suppose we are given a state $\boldsymbol{\rho} = \sum_x p_x \boldsymbol{\rho}_x$ and we wish to determine the random variable X with values x. We must do this by performing quantum measurements; any such measurement is described by a POVM $\{\mathcal{M}_y\}$ labelled by a variable Y with outcomes y. Here $p(y|x) = \text{tr}\boldsymbol{\rho}_x\mathcal{M}_y$ defines a classical channel. The Holevo bound constrains how much information can be transmitted between the two classical random variables X and Y:

$$\text{Holevo bound:} \qquad I(X:Y) \leq \chi(p_x, \boldsymbol{\rho}_x) \quad . \tag{4.18}$$

Lemma: The Holevo quantity is monotonic: $\chi(p_i, \mathcal{E}(\boldsymbol{\rho}_i)) \leq \chi(p_i, \boldsymbol{\rho}_i)$. A proof[30] follows from the observation we essentially made already around (4.9) when we introduced the state $\boldsymbol{\rho}_{AB} \equiv \sum_x p_x \boldsymbol{\rho}_x \otimes |x\rangle\langle x|$ with an extra register B that records x. The Holevo quantity for a distribution of density matrices $\boldsymbol{\rho}_x$ on A can be written as a mutual information (and hence a relative entropy):

$$\chi(p_x, \boldsymbol{\rho}_x) = I(A:B) = D\left(\boldsymbol{\rho}_{AB}||\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B\right) \ .$$

Then monotonocity of the relative entropy under quantum channels immediately shows that quantum channels cannot increase the Holevo quantity. ■

Why does the lemma imply the Holevo bound? Because we can regard the measurement $\{\mathcal{M}_y\}$ as a special case of a quantum channel $A \to Y$:

$$\boldsymbol{\rho} \mapsto \mathcal{M}(\boldsymbol{\rho}) \equiv \sum_y \left(\text{tr}\boldsymbol{\rho}\mathcal{M}_y\right)|y\rangle\langle y| \equiv \sum_y p_y |y\rangle\langle y|$$

where $\mathcal{H}_Y$ is a register which records the outcome on orthonormal states $|y\rangle$. (Complete positivity follows from $\mathcal{M}_x \geq 0$ and trace-preserving follows from $\sum_x \mathcal{M}_x = \mathbb{1}$.) Now monotonicity of the Holevo quantity says

$$\chi(p_x, \boldsymbol{\rho}_x) \geq \chi\left(p_x, \mathcal{M}(\boldsymbol{\rho}_x)\right).$$

[30] more linear than the one in C&N §12.1.1 on which Alice and Bob intrude unnecessarily; I learned it from this nice paper by Ruskai, which also contains two other proofs of this statement and various generalizations.

The RHS here unpacks exactly to $I(X:Y)$, when we identify $p(y|x) = \mathrm{tr}\boldsymbol{\rho}_x\mathcal{M}_y$:

$$\begin{aligned}\chi(p_x, \boldsymbol{\rho}_x) &\geq S\left(\mathcal{M}\left(\boldsymbol{\rho}\right)\right) - \sum_x p_x S\left(\mathcal{M}(\boldsymbol{\rho}_x)\right) \\ &= S\left(\sum_{xy} p_x \mathrm{tr}\boldsymbol{\rho}_x\mathcal{M}_y \left|y\right\rangle\left\langle y\right|\right) - \sum_x p_x S\left(\sum_y \mathrm{tr}\boldsymbol{\rho}_x\mathcal{M}_y \left|y\right\rangle\left\langle y\right|\right) \\ &= \underbrace{S\left(\underbrace{\sum_{xy} p_x p(y|x)}_{=\sum_y p(y)} \left|y\right\rangle\left\langle y\right|\right)}_{=H(Y)} - \sum_x p_x \underbrace{S\left(\sum_y p(y|x)\left|y\right\rangle\left\langle y\right|\right)}_{=H(Y|X=x)} \\ &= H(Y) - \sum_x p_x H(Y|X=x) = H(Y) - H(Y|X) = I(X:Y). \qquad \boxed{4.18}\end{aligned}$$

The Holevo bound is a sharpening of the concavity of the entropy (4.8), which showed merely that χ was positive. So we now know:

$$I(X:Y) \overset{\text{Holevo}}{\leq} \chi\left(\{p_x, \boldsymbol{\rho}_x\}\right) \overset{(4.12)}{\leq} H(X)\ .$$

This bound constrains the amount of classical information we can send with a quantum channel. Perhaps more usefully, the information about the state $\boldsymbol{\rho}$ we can extract by a POVM (into a classical RV Y) in this way is called *accessible information.* The above bound holds for any POVM. Which is the best one to use to extract all of the accessible information? I think this is a hard question in general.

We saw that (4.12) was saturated when the $\boldsymbol{\rho}_x$ were supported on orthogonal subspaces. If this is not the case, then there's no choice of POVM from which we can completely determine the distribution for X. It isn't too surprising that we can't perfectly distinguish non-orthogonal states. Only in the case where the Holevo quantity is totally squeezed on both sides, $I(X:Y) = H(X)$, so that $H(X|Y) = 0$, can we determine X completely from our knowledge of Y.

Outline of proof of monotonicity of relative entropy:

0) **Lieb's Theorem.** Consider any matrix $\mathbf{X}$ and $s \in [0,1]$. The function

$$(\mathbf{A}, \mathbf{B}) \mapsto f_{s,\mathbf{X}}(\mathbf{A}, \mathbf{B}) \equiv \mathrm{tr}\mathbf{X}^\dagger\mathbf{A}^{1-s}\mathbf{X}\mathbf{B}^s$$

is *jointly concave* in $(\mathbf{A}, \mathbf{B})$. Jointly concave means

$$f\left(\sum_i p_i\mathbf{A}_i, \sum_i p_i\mathbf{B}_i\right) \geq \sum_i p_i f(\mathbf{A}_i, \mathbf{B}_i).$$

Jointly concave is a stronger condition than concave in each argument separately, though it's not so easy to find a function which shows this.

There is an elementary proof of Lieb's theorem in Appendix 6 of C&N (it is due to Barry Simon I believe). It is satisfying (but perhaps in a similar way that programming in assembly language can be) and I've been debating whether to discuss it. But I think our time is more usefully spent in other ways. Let me know if you disagree and I am happy to talk about it.

1) **Lieb's Theorem implies joint convexity of the relative entropy.** In particular it says that for any two density matrices, the following is jointly concave in $\boldsymbol{\rho}, \boldsymbol{\sigma}$:

$$\partial_s f_{s,\mathbb{1}}(\boldsymbol{\rho}, \boldsymbol{\sigma})|_{s=0} = \lim_{s\to 0} \frac{f_{s,\mathbb{1}}(\boldsymbol{\rho}, \boldsymbol{\sigma}) - f_{0,\mathbb{1}}(\boldsymbol{\rho}, \boldsymbol{\sigma})}{s} = \lim_{s\to 0} \frac{\mathrm{tr}\boldsymbol{\rho}^{1-s}\boldsymbol{\sigma}^s - \mathrm{tr}\boldsymbol{\rho}}{s}.$$

Using $\mathrm{tr}\boldsymbol{\rho}^{1-s}\boldsymbol{\sigma}^s = \mathrm{tr}\boldsymbol{\rho}e^{-s\log\boldsymbol{\rho}}e^{s\log\boldsymbol{\sigma}} = \mathrm{tr}\boldsymbol{\rho}\,(1 - s\log\boldsymbol{\rho} + ...)\,(1 + s\log\boldsymbol{\sigma} + ...) = \mathrm{tr}\boldsymbol{\rho} - sD(\boldsymbol{\rho}||\boldsymbol{\sigma}) + \mathcal{O}(s)$, we have $\partial_s f_{s,\mathbb{1}}(\boldsymbol{\rho}, \boldsymbol{\sigma})|_{s=0} = -D(\boldsymbol{\rho}||\boldsymbol{\sigma})$. ■ [End of Lecture 12]

2) **Joint convexity of the relative entropy implies monotonicity of the relative entropy.** If you think of the sum in the partial trace $\mathrm{tr}_B\boldsymbol{\rho}_{AB}$ as a convex decomposition of $\boldsymbol{\rho}_B$ then maybe this follows immediately. More carefully, though, we can use the following result (which is an exercise in C&N):

Lemma: any matrix $\mathbf{A}$ can be *scrambled*, *i.e.* there exists a collection of unitaries $\mathbf{U}_a$ so that

$$\sum_a p_a \mathbf{U}_a \mathbf{A} \mathbf{U}_a^\dagger = \frac{1}{d}\mathrm{tr}\mathbf{A}\mathbb{1}$$

where the set of $\mathbf{U}_a$s can be chosen independent of $\mathbf{A}$, and $\sum_a p_a = 1, p_a \geq 0$. Proof of lemma: Suppose $\mathbf{A}$ is $d \times d$. Regard the space of matrices End($\mathcal{H}$) as a vector space over $\mathbb{C}$ with the Hilbert-Schmidt norm $\langle A, B\rangle = \mathrm{tr}A^\dagger B$. We can find an orthogonal basis for this space (over $\mathbb{C}$) using d^2 unitary matrices $\mathbf{U}_a$:

$$\mathrm{tr}\mathbf{U}_a^\dagger \mathbf{U}_b = \delta_{ab} d.$$

The completeness relation for this basis implies the desired relation, for any $\mathbf{A}$, with $p_a = 1/d^2$.[31][32] ■

[31] In the case where $\mathbf{A}$ is hermitian it is possible to do this scrambling with fewer (only d) matrices. Thanks for Wei-ting Kuo for showing me how to do it.

[32] In the opposite direction, a more overkill method is to use the Haar measure on $\mathsf{U}(d)$, which has a completeness relation

$$\int d\Omega(U) U_{ij} U_{kl}^\dagger = \delta_{jk}\delta_{il}$$

which implies $\int d\Omega(U) U A U^\dagger = \mathrm{tr}A\mathbb{1}$.

Once you believe this, then we can apply it to the matrix elements in A of the joint density matrix $\sum_a \mathbf{U}_a \left(\langle j| \boldsymbol{\rho}_{AB} |i\rangle\right) \mathbf{U}_a^\dagger$ – regard this as a collection of operators on B whose trace is $\boldsymbol{\rho}_A$. Use the previous result for all $|i\rangle, |j\rangle \in \mathcal{H}_A$– it is important that the $\mathbf{U}$s don't depend on ij. Then:

$$\sum_a p_a \mathbf{U}_a \boldsymbol{\rho}_{AB} \mathbf{U}_a^\dagger = \boldsymbol{\rho}_A \otimes 1\!\!1_B/|B|.$$

Then plug this into joint convexity of the relative entropy:

$$\begin{aligned} D(\boldsymbol{\rho}_A||\boldsymbol{\sigma}_A) = D(\boldsymbol{\rho}_A \otimes 1\!\!1_B/d||\boldsymbol{\sigma}_A \otimes 1\!\!1_B/d) &\leq \sum_a p_a D(\mathbf{U}_a \boldsymbol{\rho}_{AB} \mathbf{U}_a^\dagger||\mathbf{U}_a \boldsymbol{\sigma}_{AB} \mathbf{U}_a^\dagger) \qquad (4.19)\\ &= \sum_a p_a D(\boldsymbol{\rho}_{AB}||\boldsymbol{\sigma}_{AB}) = D(\boldsymbol{\rho}_{AB}||\boldsymbol{\sigma}_{AB}) \end{aligned}$$

where at the penultimate step we used the basis independence of the relative entropy. ■

On the homework you can show the converse: monotonicity implies joint convexity.

Alternate route to SSA. [Petz' book] The exponential of a self-adjoint operator is positive,

$$\exp\left(\log \boldsymbol{\rho}_{AB} - \log \boldsymbol{\rho}_B + \log \boldsymbol{\rho}_{BC}\right) = \lambda \boldsymbol{\omega}$$

and hence proportional to a density operator $\boldsymbol{\omega}$. (Notice that this is not the same as $\boldsymbol{\rho}_{AB}\boldsymbol{\rho}_B^{-1}\boldsymbol{\rho}_{BC}$ which is not necessarily even Hermitian, since the marginals don't necessarily commute.) But then

$$\begin{aligned} S(\boldsymbol{\rho}_{AB}) + S(\boldsymbol{\rho}_{BC}) - S(\boldsymbol{\rho}_{ABC}) - S(\boldsymbol{\rho}_B) &= \mathrm{tr}\boldsymbol{\rho}_{ABC}\left(\log \boldsymbol{\rho}_{ABC} - (\log \boldsymbol{\rho}_{AB} - \log \boldsymbol{\rho}_B + \log \boldsymbol{\rho}_{BC})\right)\\ &= D(\boldsymbol{\rho}_{ABC}||\lambda\boldsymbol{\omega}) = \underbrace{D(\boldsymbol{\rho}_{ABC}||\boldsymbol{\omega})}_{\geq 0} - \log \lambda \qquad (4.20) \end{aligned}$$

which implies SSA *if* we can show that $\lambda \leq 1$. It looks so innocent!

We have

$$\lambda = \mathrm{tr}\exp\left(\underbrace{\log \boldsymbol{\rho}_{AB}}_{\equiv S} + \underbrace{-\log \boldsymbol{\rho}_B}_{\equiv R} + \underbrace{\log \boldsymbol{\rho}_{BC}}_{\equiv T}\right) \qquad (4.21)$$

and would like to show that this is ≤ 1. The *Golden-Thompson* inequality says that for any two self-adjoint operators R, S,

$$\mathrm{tr}e^{R+S} \leq \mathrm{tr}e^R e^S.$$

You might be tempted to just stick a third one in there, but it's not true that $\mathrm{tr}e^{R+S+T} \overset{?}{\leq} \mathrm{tr}e^R e^S e^T$. To see a path forward, notice the following interesting formula for the inverse of a self-adjoint operator:

$$X^{-1} = \int_0^\infty dt \, (t\mathbb{1} + X)^{-2} \,.$$

Prove it by using the spectral decomposition. Lieb showed that distributing the factors differently inside the trace gives a correct inequality[33]:

$$\mathrm{tr}e^{R+S+T} \leq \int_0^\infty dt \mathrm{tr} \left(\left(t\mathbb{1} + e^{-R}\right)^{-1} e^S \left(t\mathbb{1} + e^{-R}\right)^{-1} e^T \right) \,.$$

For all this to work, we define all inverse to act as the identity on the kernel.

Now here comes the magic. Applying this to (4.21), the traces over A and C turn everything into $\boldsymbol{\rho}_B$s (which commutes with itself):

$$\begin{aligned} \lambda \quad & \leq \quad \int_0^\infty dt \; \mathrm{tr}_{ABC} \boldsymbol{\rho}_{AB} \left(t\mathbb{1} + \boldsymbol{\rho}_B\right)^{-1} \boldsymbol{\rho}_{BC} \left(t\mathbb{1} + \boldsymbol{\rho}_B\right)^{-1} \\ & = \quad \int_0^\infty dt \; \mathrm{tr}_B \left(\mathrm{tr}_A \boldsymbol{\rho}_{AB}\right) \left(t\mathbb{1} + \boldsymbol{\rho}_B\right)^{-1} \left(\mathrm{tr}_C \boldsymbol{\rho}_{BC}\right) \left(t\mathbb{1} + \boldsymbol{\rho}_B\right)^{-1} \\ & \overset{\boldsymbol{\rho}_B = \sum_b p_b |b\rangle\langle b|}{=} \quad \sum_b p_b^2 \underbrace{\int_0^\infty dt \left(\frac{1}{t + p_b}\right)^2}_{=1/p_b} = \sum_b p_b = 1. \end{aligned} \tag{4.22}$$

This proof has the advantage of giving a condition for saturating SSA, namely:

$$I(A:C|B) = 0 \quad \Leftrightarrow \quad \log \boldsymbol{\rho}_{ABC} = \log \boldsymbol{\rho}_{AB} - \log \boldsymbol{\rho}_B + \log \boldsymbol{\rho}_{BC},$$

which is visibly a quantum version of (the log of) the Markov chain equation following from $H(A:C|B) = 0$:

$$H(A:C|B) = 0 \quad \Leftrightarrow \quad p(abc) = \frac{p(ab)p(bc)}{p(b)}.$$

There is much more to say about this; if you are impatient see Ruskai, Hayden et al.

[33] A proof of this statement [see again this paper] follows from:

- For self-adjoint K and $A > 0$, the function $F(A) = \mathrm{tr}e^{K + \log A}$ is concave in A. This follows from Lieb's theorem quoted above, but apparently not in a simple way.
- The operator identity

  $$\log(M + xN) - \log M = \int_0^\infty dt \, (M + t\mathbb{1})^{-1} \, xN \, (M + t\mathbb{1})^{-1}$$

  (actually we only need the small-x limit).

5 Applications of (mostly) SSA to many body physics

[The discussion of the first two results here follows Grover.]

• **Monotonicity of the entanglement entropy in subsystem size.** Consider a region of space shaped like a slab: it has width ℓ. In the other directions it extends over the whole system, for example, we could take periodic boundary conditions in those directions, with length $L_\perp$. Sprinkle qbits over the whole space to make a quantum many-body Hilbert space. Consider any state of the whole system and let $\boldsymbol{\rho}(\ell)$ be the reduced density matrix of the slab. As the notation suggests, we assume translation invariance (for the moment at least). SSA implies:

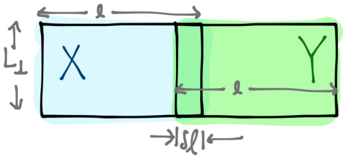
$$S(\ell) \geq S(\ell - \delta\ell) \tag{5.1}$$

that is, $\partial_\ell S(\ell) \geq 0$ (if we are allowed to take derivatives).

To see this, we use SSA in the form (the one on the homework)

$$S(X) + S(Y) \geq S(X \setminus Y) + S(Y \setminus X)$$

applied to the regions in the figure. The LHS is $2S(\ell)$ and the RHS is $2S(\ell - \delta\ell)$.

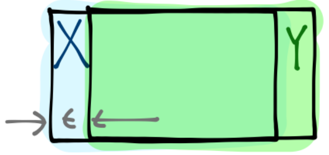


An important additional comment which I missed on the first pass (thanks to Tarun Grover for explaining this to me): we've just shown that in a translation-invariant system, the entanglement entropy of a subsystem $S(\ell)$ grows monotonically with ℓ. On the other hand, suppose the whole system is finite, of length L in the direction we've been considering (horizontal in the figure, call it x), and in pure state: you know that when $\ell \to L$, the entropy must go to zero, since $S(\ell) = S(L - \ell)$. Where is the loophole?

The loophole is that if the x-direction has period L, then when $2\ell > L$, the intersection between X and Y is not just the shaded region, but rather they must touch each other also on the other side!

• **Concavity of the entropy in subsystem size.** Along the same lines, applying SSA in the inclusion-exclusion form, with the regions at right, gives

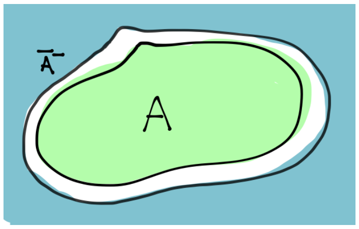
$$\begin{aligned} S(X) + S(Y) &\geq S(X \cap Y) + S(Y \cup X) \\ 2S(\ell) &\geq S(\ell + \epsilon) + S(\ell - \epsilon) \end{aligned} \tag{5.2}$$



which says that $S(\ell)$ is a concave function. If we can take $\epsilon \to 0$, it says that $\partial_\ell^2 S \leq 0$. More precisely, in a lattice model it doesn't make much sense to have ϵ less than the lattice spacing, but we must take $\epsilon \ll$ the system size and any correlation lengths.

Comments on short distance issues and the area law. You might (should) worry that I am suddenly speaking about a continuum limit, where the sites in our quantum many body system are close together compared to the lengths ℓ we are considering, so that the number of sites per unit volume is arbitrarily large. If each site is entangled with its neighbor (a finite amount), and we make an entanglement cut across arbitrarily many neighbors, we will generate a (UV divergent) entanglement entropy. No joke. This is an inevitable contribution to the entanglement entropy in a quantum many body system. It is *non-universal* in the sense that it depends on details of the arrangement of our lattice sites.

In the above (and below), we will always consider differences of entanglement entropies, in states which have the same high-energy structure. For a region of space A, let A^- denote the region A with a strip of width ξ around its boundary removed. By replacing $Y \to (Y \setminus X) \cup (X \cap Y)^-$ in the above arguments we eliminate the problem at the price of a small error. This is going to come up again.



It might be natural to talk about some of the things in §7 at this point.

Comment on translation invariance. The application Tarun Grover makes of the above inequalities is to highly disordered systems, where the couplings in $\mathbf{H}$ vary randomly in space. One is interested instead in the *average* behavior of the entanglement entropy, averaged over some ensemble of Hamiltonians. However, the above inequalities (the raw forms of SSA, before we assumed translation invariance) are true for each instance, and hence they are true of the averages as well.

• **Bounds on rates of entropy increase.** [Afkhami-Jeddi and Hartman (AJ-H). Related interesting work is this paper.] Consider a relativistic quantum field theory in d space dimensions, and consider a region of space A. Let the reduced density matrix (any state) of the subregion be $\boldsymbol{\rho}_A$. Let ρ_A^T be a thermal state with T chosen so that it has the same energy density as $\boldsymbol{\rho}_A$, *i.e.* $\mathrm{tr}\boldsymbol{\rho}\mathbf{H} = \mathrm{tr}\boldsymbol{\rho}_T\mathbf{H}$. The reduced state of the thermal state is approximately thermal: that is,

$$\boldsymbol{\rho}_A^T \simeq \frac{e^{-H_T^{(A)}}}{\mathrm{tr}_A e^{-H_T^{(A)}}} \tag{5.3}$$

where $H^{(A)}$ is just the terms in the Hamiltonian which act on the subsystem A. The approximation in (5.3) is in ignoring the terms near the boundary and their effects; in the limit of large region A, we can ignore them. (Large region A means $V_A \gg \xi^d$, large compared to the correlation length $\xi \sim 1/T$.)

[End of Lecture 13]

As in our proof that the thermal state is the maximum entropy state with the right energy, consider relative entropy

$$D(\boldsymbol{\rho}_A||\boldsymbol{\rho}_A^T) = \text{tr}\left(\boldsymbol{\rho}_A \log \boldsymbol{\rho}_A - \boldsymbol{\rho}_A \log \boldsymbol{\rho}_A^T\right) = S(\boldsymbol{\rho}_A^T) - S(\boldsymbol{\rho}_A) + \underbrace{\left\langle \beta H^{(A)} \right\rangle - \left\langle \beta H^{(A)} \right\rangle_T}_{=0}$$

where the terms which are canceling are the expectations of the energy in the state $\boldsymbol{\rho}_A$ and the thermal state. The first term is the thermal entropy, which is extensive: $S(\boldsymbol{\rho}_A^T) = V_A s_T + S_\epsilon(A)$ where s_T is the thermal entropy density, V_A is the volume of A (for more on the extensivity and the existence of s_T see the next point), and S_ϵ is the sub-extensive short-distance temperature-independent junk, which is the same as in $S(\boldsymbol{\rho}_A) \equiv S_\epsilon(A) + \hat{S}(\boldsymbol{\rho}_A)$. This leaves

$$D(\boldsymbol{\rho}_A||\boldsymbol{\rho}_A^T) = s_T V_A - \hat{S}(\boldsymbol{\rho}_A).$$

Now let us apply monotonicity of the relative entropy. First, if we consider a region $B \subset A$ completely contained in A, tracing out $B \setminus A$ gives

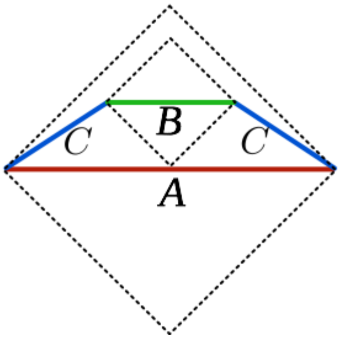
$$D(\boldsymbol{\rho}_A||\boldsymbol{\rho}_A^T) \geq D(\boldsymbol{\rho}_B||\boldsymbol{\rho}_B^T)$$

and hence

$$\hat{S}_A - \hat{S}_B \leq s_T \left(V_A - V_B\right) \equiv s_T \Delta V. \tag{5.4}$$

This gives an upper bound on $\hat{S}_A$, and on how different the entropy of A can be from that of a region inside it. (You can get a bound on how much it can shrink from SSA in the form (5.2).)

To get a bound on rate of entropy change in time, first we note that in a relativistic theory, Poincaré transformations are realized as unitary operators; this means that the states of regions A and BC in the figure at right – which are (locally) two different time slicings – are related by a unitary, and hence those of A and B are related by a quantum channel (which forgets C). That is:



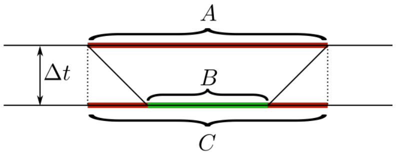
$$D(\boldsymbol{\rho}_B||\boldsymbol{\rho}_B^T) \overset{\text{MRE}}{\leq} D(\boldsymbol{\rho}_{BC}||\boldsymbol{\rho}_{BC}^T) \overset{\text{Lorentz}}{=} D(\boldsymbol{\rho}_A||\boldsymbol{\rho}_A^T) \ .$$

The idea is that A is a Cauchy surface which determines the state on the slice BC – all of the information required to know the state at BC is there at A (and vice versa). More generally, in a relativistic field theory, there is a unitary operator relating states on any two slicings of a *causal diamond*, so the relative entropy only depends on the diamond, not on the slicing. [34] Notice that it is *not* true that the state of A is related

[34] For more on this point, a good place to start is §2 of this paper.

by a unitary to the state of A at a later time – in that case, information from $\bar{A}$ can reach parts of A, so $\boldsymbol{\rho}_A$ itself evolves by open-system evolution. But Lorentz invariance forbids anything outside A from influencing the state on the slice BC (or anything else in the causal diamond of A) – whatever initial entanglement A shares with its complement remains in the state of BC.

Now consider a slab again (with regions relabelled as in the figure at right, so that now B is the region whose causal future at time dt is A, and C is just the time translation of A), and consider time evolution by an infinitesimal step dt.



[This and the previous fig are from [AJ-H]]

$$\hat{S}_A \overset{(5.4)}{\leq} \hat{S}_B + s_T(V_A - V_B) \overset{(5.1)}{\leq} \hat{S}_C + s_T(V_A - V_B)$$

from which we conclude (using $\hat{S}_A - \hat{S}_C = \Delta t \partial_t \hat{S}(\ell, t)$ and $V_A - V_B = 2c\Delta t L_\perp^{d-1}$)

$$|\partial_t \hat{S}(\ell, t)| \leq 2cL_\perp^{d-1} s_T.$$

(The bound for the rate of decrease comes from same picture with time going the other way.)

The second step seems rather conservative and perhaps a tighter bound is possible. The important thing about the slab geometry for the previous calculation was the fact that we knew that the entropy was monotonic in the slab width. The paper linked above argues that this bound generalizes to convex regions in the form $|\partial_t \hat{S}_A(t)| \leq c s_T \text{Area}(\partial A)$.

This is a version of the Small Incremental Entangling statement, about which more below in §7.3.

• **1st law of Entanglement Thermodynamics.** [following Blanco-Casini-Hung-Myers] Given any density matrix, its logarithm is a hermitian operator:

$$\boldsymbol{\rho} \equiv \frac{e^{-\mathbf{K}}}{\text{tr} e^{-\mathbf{K}}}.$$

(The additive normalization of $\mathbf{K}$ is chosen to make $\text{tr}\boldsymbol{\rho} = 1$ manifest.) $\mathbf{K}$ is called the *modular Hamiltonian* (by axiomatic field theorists) or *entanglement Hamiltonian* (by condensed matter theorists). It is generically not a sum of local operators, even if $\boldsymbol{\rho}$ is a reduced density matrix in the groundstate of a local Hamiltonian.

For some special models with extra symmetry $\mathbf{K}$ is of a known form and is local. Two examples are: (1) for a relativistic QFT in its vacuum state, the entanglement Hamiltonian for a half-space is the generator of boosts.[35] (2) for a conformal field

[35] This result is due to Bisognano and Wichmann and was rediscovered by Unruh and Weiss in

theory in the vacuum state, the entanglement Hamiltonian for a round ball can also be written in terms of an integral of the stress-energy tensor.

For a thermal state, $\mathbf{K} = \mathbf{H}$. For a reduced density matrix of a region A of size much larger than the the correlation length, when the whole system is in a thermal state, we just argued that $\mathbf{K} \approx \mathbf{H}$. [36]

Consider the relative entropy of any two states:

$$0 \leq D(\boldsymbol{\rho}_1 || \boldsymbol{\rho}_0) = \text{tr}\boldsymbol{\rho}_1 \mathbf{K}_1 - \text{tr}\boldsymbol{\rho}_0 \mathbf{K}_1 - S(\boldsymbol{\rho}_1) + S(\boldsymbol{\rho}_0) \equiv \Delta \langle \mathbf{K}_1 \rangle - \Delta S.$$

This gives a bound on the entropy difference:

$$\Delta S \leq \Delta \langle \mathbf{K}_1 \rangle .$$

This statement isn't so useful if you don't know $\mathbf{K}_1$. But now consider a smoothly-varying family of states $\boldsymbol{\rho}_\lambda$, with $\lambda \in (-\epsilon, 1]$. The function

$$f(\lambda) \equiv D(\boldsymbol{\rho}_\lambda || \boldsymbol{\rho}_0) = \underbrace{D(\boldsymbol{\rho}_0 || \boldsymbol{\rho}_0)}_{=0} + \lambda \partial_\lambda D(\boldsymbol{\rho}_\lambda || \boldsymbol{\rho}_0) + ...$$

can't be linear near $\lambda = 0$ because $D(\cdot || \cdot) \geq 0$. Therefore:

$$0 = \partial_\lambda D(\boldsymbol{\rho}_\lambda || \boldsymbol{\rho}_0)|_{\lambda=0} = \delta \langle \mathbf{K} \rangle - \delta S.$$

This is just like the first law $0 = dE - TdS$ for nearby thermodynamic equilibria.

Monotonicity of the relative entropy also implies

$$0 \leq \partial_R D(\boldsymbol{\rho}_1 || \boldsymbol{\rho}_0) = \partial_R \left(\Delta \langle \mathbf{K}_1 \rangle - \Delta S \right)$$

where R is the size of the region in question.

- **Extensivity of the entropy.** [Wehrl review, page 248] SSA can be used to argue that the *entropy density*

$$s \equiv \lim_{V \to \infty} \frac{S(V)}{|V|} \tag{5.5}$$

studies of the experience of an accelerating particle detector in QFT. I recommend this reference as a starting point.

[36] But the expectation that $\mathbf{K} \approx \mathbf{H}$ is much more general. In particular, if $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ is the reduced density matrix of a subsystem A when the whole system is in any state of finite energy density (for example a pure energy eigenstate with E/V finite), and A is a small enough fraction of the whole system, this expectation is called the *eigenstate thermalization hypothesis.* The restriction on the size of A is so that $\bar{A}$ is big enough to play the role of a heat bath for A. The idea is just as in the derivation of the canonical ensemble from the microcanonical ensemble. As appealing as this statement is, it is however frustratingly difficult to support analytically: finely tuned, integrable systems, which we can solve, can violate it. (Integrable systems which we can't solve can also violate it; that's called *many-body localization.*) I strongly recommend this paper for evidence and further references, and estimates of how surprisingly big A can be.

exists (it might be zero) in translation-invariant systems in the thermodynamic limit. It uses the same trick as above of intersecting translates of a given region.

Briefly, consider again a slab geometry. Subadditivity $S(\ell_1 + \ell_2) \leq S(\ell_1) + S(\ell_2)$ is not quite enough to guarantee that the limit above exists. No discussion of analysis would be complete without a horrifying and unphysical counterexample involving the rational numbers, so here we go: Consider the function defined on the set of intervals of the real line $\mathfrak{Q}([a,b]) = \begin{cases} 0, & b-a \in \mathbb{Q} \\ \infty, & \text{else} \end{cases}$. (Argh.) This function is subadditive, but the limit defined in (5.5) certainly does not exist.

Anyway, SSA rules out this counterexample (and all others) by placing a bound on S. For a subadditive and bounded function, the limit in (5.5) exists (according to an assembly-language theorem of Szego and Polya). How does SSA place a bound on $S(\ell)$? Make a slab of length ℓ by intersecting two slabs of length $\ell_0 > \ell$ called X and Y. Then $S(X \cap Y) + S(X \cup Y) \leq S(X) + S(Y)$ says

$$S(\ell) + \underbrace{S(2\ell_0 - \ell)}_{\geq 0} \leq 2S(\ell_0) \quad \implies \quad S(\ell) < 2S(\ell_0).$$

So this shows that, at least for slab-like regions, the thermal entropy of translation invariant states can't be super-extensive, even in the continuum limit.

6 Entanglement as a resource

6.1 When is a mixed state entangled?

I need to fill a hole in the above discussion: Above we said that a pure bipartite state $|w\rangle$ is an entangled on AB when the Schmidt rank is larger than one. The Schmidt decomposition is something we know how to do for pure states. What does it mean for a mixed state on AB to be entangled or not? We answer by saying when it is not:

For vividness imagine that A and B are separated by a big distance. Surely you agree that $\boldsymbol{\rho} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$ is not entangled. But now suppose that A flips a coin and as a result does some unitary $\mathbf{U}_a^A \otimes \mathbb{1}_B$ with probability p_a to her state:

$$\boldsymbol{\rho} \to \sum_a p_a \left(\mathbf{U}_a^A \boldsymbol{\rho}_A \left(\mathbf{U}_a^A\right)^\dagger \right) \otimes \boldsymbol{\rho}_B.$$

Even better, A picks up the telephone and tells B the result of the coin flip, and so B does some unitary $\mathbb{1}_A \otimes \mathbf{U}_a^B$ to his state:

$$\boldsymbol{\rho} \to \sum_a p_a \left(\mathbf{U}_a^A \boldsymbol{\rho}_A \left(\mathbf{U}_a^A\right)^\dagger \right) \otimes \left(\mathbf{U}_a^B \boldsymbol{\rho}_B \left(\mathbf{U}_a^B\right)^\dagger \right). \tag{6.1}$$

These operations are called *local operations* (unitaries which act as $\mathbf{U}^A \otimes \mathbb{1}_B$) and *classical communication* (the telephone), or altogether: LOCC. Mixed states of the form (6.1) are not entangled (sometimes called *separable* but not by me).

Examples where we have seen LOCC in action are the quantum teleportation algorithms (on the homework).

6.2 States related by LOCC

[C&N §12.5] The problem of when is a density matrix factorizable by LOCC is a special case of a more general question: which states are related by this LOCC operation

$$\boldsymbol{\rho} \overset{\text{LOCC}}{\mapsto} \sum_a p_a \left(\mathbf{U}_a^A \otimes \mathbf{U}_a^B\right) \boldsymbol{\rho} \left(\mathbf{U}_a^A \otimes \mathbf{U}_a^B\right)^\dagger \quad ? \tag{6.2}$$

Notice what the LOCC operation (6.2) does to the reduced density matrix on A:

$$\boldsymbol{\rho}_A \overset{\text{LOCC}}{\mapsto} \sum_a p_a \mathbf{U}_a^A \boldsymbol{\rho}_A \left(\mathbf{U}_a^A\right)^\dagger = \mathcal{E}(\boldsymbol{\rho}_A)$$

– it's a quantum expander. As we'll see in more detail (and as you saw on the homework) this is not an equivalence relation, since it's not reflexive.

Majorization. A fact about the action of quantum expanders is relevant here: the output of such a channel $\boldsymbol{\rho} = \mathcal{E}(\boldsymbol{\sigma})$ *majorizes* the input. This means that if we order their eigenvalues $\{\rho_i^\downarrow\}$ and $\{\sigma_i^\downarrow\}$ in decreasing order (indicated by the superscript downarrow), then

$$\text{for all } n, \ \sum_{i=1}^n \rho_i^\downarrow \le \sum_{i=1}^n \sigma_i^\downarrow, \quad \Leftrightarrow \quad \boldsymbol{\rho} \prec \boldsymbol{\sigma}.$$

(Since we are interested in probabilities and density matrices, equality must hold for $n = \dim \mathcal{H}$.) This means that the output is *more mixed* than the input, as quantified for example by the purity $\text{tr}\boldsymbol{\rho}^2 = \sum_i \rho_i^2 \le \sum_i \sigma_i^2 = \text{tr}\boldsymbol{\sigma}^2$, or indeed for any convex function f, $\text{tr} f(\boldsymbol{\rho}) \le \text{tr} f(\boldsymbol{\sigma})$ (or by the von Neumann entropy which should *increase* because it is concave).

This is a partial order on the space of density matrices (and hence probability distributions). Partial order means not every pair of distributions participates in such a relation. It is useful to pronounce the symbol $\prec$ as 'is less pure than'.

[End of Lecture 14]

For example, on a d-dimensional Hilbert space, the diagonal-part channel Φ_{QC} is a quantum expander with d unitaries $\mathbf{Z}^i, i = 1..d$, with $\mathbf{Z}$ the clock operator. The fact that its image is more mixed is the statement that the sum of the n largest diagonal entries of any hermitian matrix is smaller than the sum of its n largest eigenvalues. This is called Ky Fan's inequality[37].

There is a nice discussion of majorization and Uhlmann's theory of mixing enhancement in the review by Wehrl with more examples.

In fact, the converse is also true:

$$\text{Uhlmann's Theorem:} \qquad \boldsymbol{\rho} = \sum_a p_a \mathbf{U}_a \boldsymbol{\sigma} \mathbf{U}_a^\dagger \iff \boldsymbol{\rho} \prec \boldsymbol{\sigma}. \tag{6.3}$$

The classical version of this statement is related to Birkhoff's theorem: a probability distribution p majorizes another q ($p \prec q$) if and only if p is made from q by a convex combination of permutations. I actually cited a version of this theorem earlier when we discussed Markov chains, because this result means also that $p_i = P_{ij} q_j$ where P is a doubly stochastic matrix[38].

$\boxed{\Longleftarrow}$ So for two density matrices related by $\boldsymbol{\rho} \prec \boldsymbol{\sigma}$, their eigenvalues satisfy $\{\rho\} \prec \{\sigma\}$ as classical distributions and hence are related by a doubly stochastic matrix (convex combination of permutations)

$$\rho_i = \sum_{a,j} p_a \pi^a_{ij} \sigma_j .$$

[37] Here's a proof: Let $|\phi_i\rangle$ be an arbitrary ON basis. Let $|i\rangle$ be the eigenbasis of ρ. Then the partial sums of the diagonal elements in this basis $\sum_{i=1}^n \langle \phi_i | \rho | \phi_i \rangle = \sum_i \langle \psi_i | \rho | \psi_i \rangle$ are independent of basis transformations within the subspace spanned by the elements with the largest diagonal matrix elements. So let's choose another basis of span$\{\phi_1 \cdots \phi_n\}$ by

$$|\psi_n\rangle \perp |1\rangle .. |n-1\rangle , \quad |\psi_{n-1}\rangle \perp |1\rangle \cdots |n-1\rangle , |\psi_n\rangle , \quad |\psi_1\rangle \perp |\psi_2\rangle , .. |\psi_n\rangle .$$

Then we have $\langle \psi_i | \rho | \psi_i \rangle \leq \lambda_i$ for sure.

[38] Another aspect of this classical analogy worth mentioning is the classical analog of the Stinespring dilation theorem: Any convex combination of permutations can be written as a single permutation on an enlarged distribution: $p_{in} \to p_{out} = \sum_a p_a \pi_a(p_{in})$ can be accomplished by

$$p_{in} \mapsto \Pi(p_{in} \otimes u) \overset{\text{partial trace}}{\mapsto} \sum_{j=1}^{\Omega} \Pi(p_{in} \otimes u)_{ij} = p_{out}(i)$$

where $u = (\frac{1}{\Omega}, \cdots \frac{1}{\Omega})$ is a uniform distribution on an auxiliary space, and Π is a permutation (the analog of unitary) on the enlarged sample space. So for example take $p_{in} = (1,0)$ to be a 'pure state'. Then $p_{in} \otimes u = (\frac{1}{\Omega}, \cdots \frac{1}{\Omega}, 0, \cdots 0)$, and $\Pi(p_{in} \otimes u)$ is a distribution half of whose entries are $\frac{1}{\Omega}$, and the other half are zero. The partial trace then gives $p_{out} = (\frac{m}{\Omega}, \frac{\Omega - m}{\Omega})$, where m is the (integer) number of nonzero entries in the first half. In this way we can take $(1,0) \mapsto (q_1, q_2)$ for any rational q_i.

[39] But the actual density matrices are

$$\boldsymbol{\rho} = \mathbf{W}\Lambda_\rho \mathbf{W}^\dagger, \quad \boldsymbol{\sigma} = \mathbf{V}\Lambda_\sigma \mathbf{V}^\dagger$$

where

$$\Lambda_\rho = \sum_a p_a \pi^a \Lambda_\sigma \left(\pi^a\right)^t = \sum_a p_a \pi^a \Lambda_\sigma \left(\pi^a\right)^\dagger$$

is the diagonal matrix with entries ρ_i (in descending order). So we have

$$\boldsymbol{\rho} = \sum_a p_a \mathbf{W}\pi_a \Lambda_\sigma \pi_a^\dagger \mathbf{W}^\dagger = \sum_a p_a \underbrace{\mathbf{W}\pi_a \mathbf{V}^\dagger}_{\equiv \mathbf{U}_a} \boldsymbol{\sigma} \underbrace{\mathbf{V}\pi_a^\dagger \mathbf{W}^\dagger}_{\equiv \mathbf{U}_a^\dagger}.$$

$\boxed{\Longrightarrow}$ If we have two density matrices related by a quantum expander, then their diagonal matrices of eigenvalues are related by $\Lambda_\rho = \sum_a p_a \mathbf{V}_a \Lambda_\sigma \mathbf{V}_a^\dagger$ which since Λ_σ is diagonal says

$$\rho_i = \sum_{ak} p_a V^a_{ik} \sigma_k \left(V^a\right)^\dagger_{ki} = \sum_{ak} p_a |V^a_{ik}|^2 \sigma_k$$

but $P_{ik} \equiv \sum_a p_a |V^a_{ik}|^2$ is doubly stochastic (positive and trace one on both indices) since V is unitary and $\sum_a p_a = 1$. $\boxed{6.3}$

Notice that it is not the case that every two density matrices are related by $\succ$ or $\prec$. Indeed more general quantum channels have Kraus operators which are not proportional to unitaries and destroy the ordering of the eigenvalue sums. For example, the amplitude damping channel increases the purity of the output relative to the input.

Now let's return to our discussion of states related by LOCC. You might worry that our definition of LOCC is too limited, because we only allowed A to send information to B in our discussion.

You might also worry that A can do things to the system which are not just unitary operations, such as measurements. Indeed A could measure something about the state, send the result to B, who could then make another measurement and send the result to A, etc...

The following result (Proposition 12.14 of C&N) shows that the most general outcome can be obtained by the following steps: A makes a measurement, sends the result to B, who performs a unitary accordingly.

[39] OK, now you'll want to know why is the classical Birkhoff theorem true, *i.e.* why for two distributions $x \prec y$ means that x is a convex combination of permutations of y. In outline: $\boxed{\Longleftarrow}$: $x \prec y$ is a convex condition on x. So we can ask if it is true for the extreme points, *i.e.* when $x = \pi y$. But clearly $x = \pi y$ for π any permutation means $x \prec y$ (and $y \prec x$ too) since the definition of majorization involves ordering the eigenvalues and hence undoing π. So this shows that $\sum_a p_a \pi y \prec y$. $\boxed{\Longrightarrow}$: see page 574 of C&N, or for more, Watrous lecture 13. Both of these statements are in turn equivalent to the condition that $x = Py$ where P is a doubly stochastic matrix.

Measurement operators. Earlier I described generalized measurements in terms of POVMs $\{E_a \geq 0, \sum_a E_a = \mathbb{1}\}$. A more refined description involves *measurement operators*, $\{\mathcal{M}_a\}$ in terms of which $E_a = \mathcal{M}_a^\dagger \mathcal{M}_a$. The additional information is that the $\mathcal{M}_a$ can be used as Kraus operators to determine the state after measurement: if the state is $|\psi\rangle$, the outcome is a (which happens with probability $|\mathcal{M}_a |\psi\rangle|^2$, the resulting state is proportional to $\mathcal{M}_a |\psi\rangle$. For mixed states, the probability is $\text{tr}\mathcal{M}_a \boldsymbol{\rho} \mathcal{M}_a^\dagger$, and the final state is $\mathcal{M}_a \boldsymbol{\rho} \mathcal{M}_a^\dagger$.

If B measures $\{\mathcal{M}^a \equiv \sum_{kl} \mathcal{M}^a_{kl} |k\rangle\langle l|_B\}$ on a state with Schmidt decomposition $|\psi\rangle = \sum_\ell \sqrt{\lambda_\ell} |\ell_A \ell_B\rangle$ the resulting state will be

$$|\psi_a\rangle \propto \mathcal{M}^a |\psi\rangle = \sum_{kl} \mathcal{M}^a \sqrt{\lambda_\ell} |\ell\rangle_A |k\rangle_B$$

with probability $\sum_{k\ell} \lambda_\ell |\mathcal{M}^a_{k\ell}|^2$.

Now let $\{\mathcal{N}^a \equiv \sum_{kl} \mathcal{M}^a_{kl} |k\rangle\langle l|_A\}$ be a set of measurement operators on A with the same matrix elements in the Schmidt basis for $|\psi\rangle$. If A measures $\mathcal{N}$ and gets a the resulting state is

$$|\phi_a\rangle \propto \mathcal{N}^a |\psi\rangle = \sum_{kl} \mathcal{M}^a \sqrt{\lambda_\ell} |k\rangle_A |\ell\rangle_B$$

with the same probability $\sum_{k\ell} \lambda_\ell |\mathcal{M}^a_{k\ell}|^2$. ϕ_a and ψ_a are related by interchange of the labels on A and B. In particular, they have the same Schmidt values. This means they are related by local unitaries:

$$|\varphi^a\rangle = U^a_A \otimes V^a_B |\psi^a\rangle .$$

So the result of B measuring $\{\mathcal{M}^a\}$ is the same as A measuring $\{U^a_A \mathcal{N}^a\}$, sending a and B doing V^a_B. So we can replace all B measurements by more A measurements. And many measurements by A $\{\mathcal{M}_1^{a_1}\}, \{\mathcal{M}_2^{a_2}\} \cdots$ is the same as one measurement by A of $\{\mathcal{M}_1^{a_1} \mathcal{M}_2^{a_2} \cdots\}$.

Nielsen's Theorem (Theorem 12.15 of C&N): A bipartite pure state $|\psi_1\rangle$ can be turned into $|\psi_2\rangle$ by LOCC between A and $\bar{A}$ if and only if $\boldsymbol{\rho}_1 \prec \boldsymbol{\rho}_2$, where $\boldsymbol{\rho}_\alpha \equiv \text{tr}_{\bar{A}} |\psi_\alpha\rangle \langle\psi_\alpha|$.

Sketch of $\boxed{\Longleftarrow}$: According to the Uhlmann theorem, the majorization of (1) by (2) means there exists a quantum expander on A so that $\boldsymbol{\rho}_1 = \sum_a p_a \mathbf{U}_a \boldsymbol{\rho}_2 \mathbf{U}_a^\dagger$. This can be used to build an instrument on A with measurement operators

$$\mathcal{M}_a \equiv \sqrt{p_a \boldsymbol{\rho}_2} \mathbf{U}_a^\dagger \boldsymbol{\rho}_1^{-1/2}. \tag{6.4}$$

By this I mean a POVM

$$\mathsf{E}_a \equiv \mathcal{M}_a^\dagger \mathcal{M}_a = \boldsymbol{\rho}_1^{-1/2} p_a \mathbf{U}_a \boldsymbol{\rho}_2 \mathbf{U}_a^\dagger \boldsymbol{\rho}_1^{-1/2}$$

(which satisfy $\sum_a \mathsf{E}_a = 1\!\!1_A$ by the quantum expander definition) but also an instruction that the state after the measurements are obtained by using the $\mathcal{M}_a$ as Kraus operators, so upon doing the measurement on state $|\psi\rangle$ and getting outcome a, the output state is $\propto \mathcal{M}_a |\psi_1\rangle$. (Note that this whole story takes place on the support of $\boldsymbol{\rho}_1$, so if $\boldsymbol{\rho}_1$ is not invertible, we define $\boldsymbol{\rho}_1^{-1}$ by padding with the identity on its kernel.) Let $\boldsymbol{\rho}_a$ be A's reduced density matrix when the outcome is a, in which case, by construction

$$\boldsymbol{\rho}_a \propto \mathcal{M}_a \boldsymbol{\rho}_1 \mathcal{M}_a^\dagger = p_a \boldsymbol{\rho}_2$$

which means that (upon normalization), $\boldsymbol{\rho}_a = \boldsymbol{\rho}_2$ for all a. A sends the result a to $\bar{A}$, who can then act with a unitary $\mathbf{V}_a$ on $\bar{A}$ to accomplish

$$1\!\!1_A \otimes \mathbf{V}_a \left(\mathcal{M}_a \otimes \mathbf{1}_{\bar{A}} |\psi_1\rangle\right) = |\psi_2\rangle .$$

$\boxed{\Longrightarrow}$: Suppose $|\psi_1\rangle$ can be turned into $|\psi_2\rangle$ by A performing a measurement with measurement operators $\mathcal{M}_a$ (so that $p_a = \mathrm{tr}_A \mathcal{M}_a \boldsymbol{\rho}_1 \mathcal{M}_a^\dagger$) and sending the result by post to $\bar{A}$, whose occupants conspire to perform an appropriate unitary $\mathbf{V}_a$. To obtain the associated unitaries, we basically just read the relation (6.4) in the other direction. More constructively: after A's measurement, by assumption, her state is $\boldsymbol{\rho}_2 \equiv \mathrm{tr}_{\bar{A}} |\psi_2\rangle \langle\psi_2|$ no matter the outcome of the measurement. But then for all a we must have

$$\mathcal{M}_a \boldsymbol{\rho}_1 \mathcal{M}_a^\dagger = p_a \boldsymbol{\rho}_2 \tag{6.5}$$

(the trace of this equation is the equation for the probability of outcome a). Do polar decomposition ($Z = \sqrt{ZZ^\dagger}\mathbf{V}$) on

$$\mathcal{M}_a \sqrt{\boldsymbol{\rho}_1} \overset{\text{polar}}{=} \sqrt{\mathcal{M}_a \boldsymbol{\rho}_1 \mathcal{M}_a^\dagger}\,\mathbf{V}_a \overset{(6.5)}{=} \sqrt{p_a \boldsymbol{\rho}_2}\,\mathbf{V}_a.$$

Now use $\sum_a \mathcal{M}_a^\dagger \mathcal{M}_a = 1\!\!1$ in $\left(\mathcal{M}_a \sqrt{\boldsymbol{\rho}_1}\right)^\dagger \mathcal{M}_a \sqrt{\boldsymbol{\rho}_1} = p_a \mathbf{V}_a^\dagger \boldsymbol{\rho}_2 \mathbf{V}_a$ to show that $\mathbf{V}_a$ are the desired unitaries which show that $\boldsymbol{\rho}_1 \prec \boldsymbol{\rho}_2$.

[Petz' book, p. 7 and 178, who attributes this result to Schrödinger.] Here is a nice lesson we can extract from this proof; it generalizes our statement that measurement (without looking at the answer) increases entropy. The spectral decomposition of

$$\boldsymbol{\rho} = \sum_i \rho_i |i\rangle \langle i| = \sum_a \mu_a |w_a\rangle \langle w_a| \qquad (\langle i|j\rangle = \delta_{ij})$$

majorizes any other ensemble preparation of a state: $\{\rho_i\} \succ \{\mu_a\}$. This is because we can find a unitary $\mathbf{V}$ so that

$$\sum_i V_{ai}\sqrt{\rho_i}\,|i\rangle = \sqrt{\mu_a}\,|w_a\rangle\,, \qquad \text{(in particular, } V_{ai} = \langle i|w_a\rangle\sqrt{\frac{\mu_a}{\rho_i}}\text{)} \tag{6.6}$$

and hence (take the norm of both sides of (6.6)) $\mu_a = \sum_i |V_{ai}|^2\rho_i$ and $\mu \prec \rho$. Here's the proof that V is unitary:

$$\sum_a V_{ai}V^\star_{aj} = \sum_a \langle i|w_a\rangle\,\mu_a\,\langle w_a|j\rangle\,\frac{1}{\sqrt{\rho_i\rho_j}} = \langle i|\underbrace{\sum_a \mu_a|w_a\rangle\langle w_a|}_{=\boldsymbol{\rho}}|i\rangle\,\frac{1}{\sqrt{\rho_i\rho_j}} = \delta_{ij}.$$

(Similarly for $\sum_i V_{ia}V^\star_{ib} = \delta_{ab}$.)

I should mention that we have focussed on a special case in the above discussion by considering only the case of LOCC between A and its complement $\bar{A}$, so that the two together are in a pure state. The generalization of Nielsen's result to mixed states is a longer story. I recommend the discussion in the notes by Watrous, specifically lecture 16. [End of Lecture 15]

6.3 Entanglement distillation, briefly

Earlier I drew some pictures where I represented the amount of entanglement between two subsystems by drawing a number of lines between them (*e.g.* in illustrating (4.15)) each of which represented a Bell pair shared by the subsystems. This statement can be made precise in the following way: n copies of the whole system AB in the given bipartite pure state $|\psi\rangle_{AB}$, can be converted by LOCC operations into $nS(A)$ Bell pairs. (In fact it is possible to go both ways in this asymptotic statement.) The construction uses the Uhlmann theorem.

This is another application of Shannon source coding. If the Schmidt representation of the state is

$$|\psi\rangle_{AB} = \sum_x \sqrt{p(x)}\,|x\rangle_A\,|x\rangle_B$$

then the tensor product of n copies is

$$\mathcal{H}_{AB}^{\otimes n} \ni |\psi\rangle_{AB}^{\otimes n} = \sum_{x_1\cdots x_n} \sqrt{p(x_1)\cdots p(x_n)}\,|x_1\cdots x_n\rangle_{A^{\otimes n}}\,|x_1\cdots x_n\rangle_{B^{\otimes n}}\,.$$

Shannon tells us that we can make a good approximation to $\psi^{\otimes n}$ by projecting onto the subspace of ϵ-typical sequences (and re-normalizing). This subspace has dimension less than $2^{n(H(p)+\epsilon)} = 2^{n(S(A)+\epsilon)}$, and the error from the re-normalizing goes to zero as $n \to \infty$.

Here's the protocol to convert n copies of $|\psi\rangle$ into $nS(A)$ Bell pairs by LOCC: A measures the projector onto the typical subspace, $\Pi = \sum_{x\in T} |x\rangle\langle x|$. If the state is not in the typical subspace, try again – it's ok we have many copies of the state. The resulting reduced state on A (call it $\boldsymbol{\rho}$) is in the typical subspace and its largest eigenvalue (and hence all the others) is bounded by

$$\frac{2^{-nS(A)-\epsilon}}{1-\delta}$$

where $1-\delta$ is the probability contained in the typical sequences[40]. This means that we can choose m such that $\frac{2^{-nS(A)-\epsilon}}{1-\delta} \leq 2^{-m}$ and then

$$\sum_{k=1}^{K} \rho_k^{\downarrow} \leq \sum_{k=1}^{K} 2^{-m} = K 2^{-m}.$$

That is, the eigenvalues of $\boldsymbol{\rho}$ are majorized by the vector of length 2^m whose entries are all 2^{-m} – which is the reduced density matrix on A of m Bell pairs shared between A and B, *i.e.* a state of the form $\frac{1}{2^{-m/2}} \otimes_{i=1}^{m} (|00\rangle + |11\rangle)_{A_i B_i} \otimes ...$ where ... is other products of states in A, B to make up the rest of their dimensions. Bam! Now just use the theorem above that says majorization implies LOCC is possible.

More generally, if ρ_{AB} is mixed rather than pure, we can ask how many Bell pairs per copy can be distilled from many copies of it. In that case the answer is not S_A, but generally smaller, because the entropy of ρ_{AB} can have classical contributions, which don't count entanglement between A and B, and hence aren't helpful for making Bell pairs.

Single-copy entanglement. Our beloved von Neumann entropy is the special case $S(\boldsymbol{\rho}) = \ln 2 \cdot \lim_{\alpha\to 1} S_\alpha(\boldsymbol{\rho})$ of the Renyi entropies:

$$S_\alpha(\boldsymbol{\rho}) \equiv \frac{\text{sgn}(\alpha)}{1-\alpha} \log \text{tr} \boldsymbol{\rho}^\alpha = \frac{\text{sgn}(\alpha)}{1-\alpha} \log \sum_a p_a^\alpha \ .$$

If we know these for enough α, we have complete information about the spectrum of $\boldsymbol{\rho}$ (for an N-dimensional $\mathcal{H}$, then N of them are enough). The case $\alpha = 0$ is just the

[40] Recall that a sequence is typical if its surprise is close to the average surprise: $x_1..x_n$ is ϵ-typical if $\left|\frac{1}{n}\log\frac{1}{p_{x_1..x_n}} - H(p)\right| \leq \epsilon$ which says

$$2^{-n(S+\epsilon)} \leq p(x_1..x_n) \leq 2^{-n(S-\epsilon)}.$$

rank of $\boldsymbol{\rho}$, which, if $\boldsymbol{\rho}$ is a reduced density matrix of a pure state $|\psi\rangle_{A\bar{A}}$, is the Schmidt rank of the state with respect to the bipartition $A\bar{A}$.

I mention these here because the special case of $\alpha = \infty$ gets a nice interpretation from entanglement distillation in the case where we have only a single copy of the state available.

$$S_\infty(\boldsymbol{\rho}) = \lim_{\alpha\to\infty} \frac{-1}{\alpha - 1} \log \sum_a p_a^\alpha = - \lim_{\alpha\to\infty} \frac{1}{\alpha} \log \left(p_1^\downarrow\right)^\alpha = -\log p_1^\downarrow$$

– it is just the log of the inverse of the largest eigenvalue.

Consider again the case $\boldsymbol{\rho} = \mathrm{tr}_{\bar{A}} |\psi\rangle \langle\psi|$. Suppose by LOCC can distill from $\boldsymbol{\rho}$ a maximally entangled state on $A\bar{A}$ of dimension M,

$$|\Phi_M\rangle \equiv \frac{1}{\sqrt{M}} \sum_{i=1}^{M} |ii\rangle_{A\bar{A}} .$$

The largest possible $M \equiv e^{E_1(\boldsymbol{\rho})}$ is a measure of how entangled this state of $A\bar{A}$ is; E_1 is called the *single-copy entanglement*. It is called this in contrast with the vN entropy which generally answers asymptotic questions about what happens if we have arbitrarily many copies of the state, as does the Shannon entropy.

If we can do this, then it must be the case that

$$\boldsymbol{\rho} \prec P_M/M$$

where P_M is a uniform projector onto an M-dimensional subspace of A. That is, we must have

$$\sum_{k=1}^{K} \rho_k^\downarrow \leq \sum_{k=1}^{K} \frac{1}{M} = \frac{K}{M}, \quad \forall K = 1..M \ .$$

These conditions are equivalent to $\rho_1^\downarrow \leq \frac{1}{M}$, since the eigenvalues are in decreasing order. That is, $M \leq \left(\rho_1^\downarrow\right)^{-1} = e^{S_\infty(\boldsymbol{\rho})}$ so $\max M = \left\lfloor e^{S_\infty} \right\rfloor \in \left[e^{S_\infty} - 1, e^{S_\infty}\right]$ and

$$E_1(\boldsymbol{\rho}) = \max \log M = \log \max M = \log \left(\left\lfloor \left(\rho_1^\downarrow\right)^{-1} \right\rfloor \right) \simeq -\log \rho_1^\downarrow = S_\infty(\boldsymbol{\rho}).$$

So the Renyi$_\infty$ entropy estimates the single-copy entanglement. (The more precise statement of '$\simeq$' here is $E_1(\boldsymbol{\rho}) \in \left[\log\left(e^{S_\infty} - 1\right), S_\infty\right]$.)

See this paper for a discussion of single-copy entanglement in critical spin chains.

Entanglement catalysis. I should mention that there is a zoo of protocols related to LOCC, with names like entanglement catalysis, embezzlement, ...

An example (C&N Ex. 12.21 and these notes of M. P. Mueller): The following two distributions on four items

$$p = (2/5, 2/5, 1/10, 1/10), \quad q = (1/2, 1/4, 1/4, 0).$$

do not participate in a majorization relation (since $p_1 < q_1$, but $p_1 + p_2 > q_1 + q_2$). But now let $c = (3/5, 2/5)$ be a distribution on some other two-valued variable. Then

$$p \otimes c = \left(\frac{2}{5} \cdot \frac{3}{5}, \frac{2}{5} \cdot \frac{3}{5}, \frac{1}{10} \cdot \frac{3}{5}, \frac{1}{10} \cdot \frac{3}{5}, \frac{2}{5} \cdot \frac{2}{5}, \frac{2}{5} \cdot \frac{2}{5}, \frac{1}{10} \cdot \frac{2}{5}, \frac{1}{10} \cdot \frac{2}{5}\right)$$

$$q \otimes c = \left(\frac{1}{2} \cdot \frac{3}{5}, \frac{1}{4} \cdot \frac{3}{5}, \frac{1}{4} \cdot \frac{3}{5}, 0, \frac{1}{2} \cdot \frac{2}{5}, \frac{1}{4} \cdot \frac{2}{5}, \frac{1}{4} \cdot \frac{2}{5}, 0\right)$$

do satisfy $p \otimes c \prec q \otimes c$.

Since majorization between density matrices is just a property of their eigenvalues, you can imagine that there are quantum versions of this statement (and in fact it seems to have been discovered in that context first): consider the states

$$|\sqrt{\rho}\rangle \equiv \sqrt{\frac{2}{10}}|00\rangle + \sqrt{\frac{2}{10}}|11\rangle + \sqrt{\frac{1}{10}}|22\rangle + \sqrt{\frac{1}{10}}|33\rangle\,, \quad |\sqrt{\sigma}\rangle \equiv \sqrt{\frac{1}{2}}|00\rangle + \sqrt{\frac{1}{4}}|11\rangle + \sqrt{\frac{1}{4}}|22\rangle$$

on $\mathcal{H}_A \otimes \mathcal{H}_B$ (each 4-state systems). It's not possible to intercommute $|\sqrt{\rho}\rangle$ and $|\sqrt{\sigma}\rangle$ by LOCC since the distributions p, q above are the eigenvalues of $\rho_A = \text{tr}_B |\sqrt{\rho}\rangle\langle\sqrt{\rho}|$ and $\sigma_A = \text{tr}_B |\sqrt{\sigma}\rangle\langle\sqrt{\sigma}|$ respectively. Now let $|\sqrt{c}\rangle = \sqrt{\frac{3}{5}}|\uparrow\uparrow\rangle + \sqrt{\frac{2}{5}}|\downarrow\downarrow\rangle$ be an ancillary pair of qbits shared by AB. The fact that $p \otimes c \prec q \otimes c$ then implies that $|\sqrt{\sigma}\rangle \otimes |\sqrt{c}\rangle \xrightarrow{\text{LOCC}} |\sqrt{\rho}\rangle \otimes |\sqrt{c}\rangle$ is possible! So an ancillary system can facilitate LOCC operations. c is called a *catalyst* since its presence allows a majorization relation, but it is not itself consumed by the process.

Notice that this means that p and q now participate in a partial order; the terminology is that p is *trumped by* q. This relation can be shown to be transitive by tensoring in both the catalysts involved. This paper describes a condition for the existence of a catalyst that allows $p \otimes c \prec q \otimes c$: *all* the Renyis (except $\alpha = 0$) of p must be larger than those of q and in addition $\sum_i \log p_i > \sum_i \log q_i$ is required.

6.4 Distance measures

Can two states which are close together have wildly different vN entropies? An answer to this question (a quantitative version of 'no') is called the Fannes inequality (a sharp improvement of which is the Fannes-Audenaert inequality).

But this begs the question: 'close' by what distance measure? More generally, to make any useful approximate statements about density matrices, it is necessary to be able to quantify the distance between a pair of them.

So far we've compared states using the relative entropy, which, as we saw, has some shortcomings as a distance. Two distance measures frequently used in the literature (and which are the subjects of the two parts of the definition-heavy C&N chapter 9) are the trace distance

$$T(\boldsymbol{\rho}, \boldsymbol{\sigma}) \equiv \frac{1}{2}\mathrm{tr}|\boldsymbol{\rho} - \boldsymbol{\sigma}| \equiv \frac{1}{2}\| \boldsymbol{\rho} - \boldsymbol{\sigma} \|_1$$

[41] and the fidelity

$$F(\boldsymbol{\rho}, \boldsymbol{\sigma}) \equiv \| \sqrt{\boldsymbol{\rho}}\sqrt{\boldsymbol{\sigma}} \|_1 = \mathrm{tr}\sqrt{\sqrt{\boldsymbol{\rho}}\boldsymbol{\sigma}\sqrt{\boldsymbol{\rho}}}.$$

They are both basis independent. They both have classical counterparts to which they reduce when the two operators share eigenbases. (The often-vexing factor of two in the trace distance is there so that $T \in [0, 1]$ The largest value is attained when $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ have orthogonal support.)

In our discussion of the mutual information bound on correlations in §7.2 it will be important that the trace distance bounds the relative entropy from below. And I've been trying to avoid thinking about the fidelity (though I may relent soon). So let's talk about trace distance a bit. It has many virtues, including monotonicity, continuity, convexity, all of which are not so difficult to see.

All the magic is in the innocent-looking absolute value. Decompose $\boldsymbol{\rho} - \boldsymbol{\sigma} \equiv \mathbf{Q} - \mathbf{R}$ where $\mathbf{Q}$ and $\mathbf{R}$ are positive operators with orthogonal support[42]. So $|\boldsymbol{\rho} - \boldsymbol{\sigma}| = \mathbf{Q} + \mathbf{R}$ and

$$T(\boldsymbol{\rho}, \boldsymbol{\sigma}) = \frac{1}{2}\mathrm{tr}|\boldsymbol{\rho} - \boldsymbol{\sigma}| = \frac{1}{2}\mathrm{tr}\left(\mathbf{Q} + \mathbf{R}\right) = \mathrm{tr}\mathbf{Q}$$

where the last step follows since both $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ have unit trace, so $\mathrm{tr}\mathbf{Q} - \mathrm{tr}\mathbf{R} = \mathrm{tr}\left(\mathbf{Q} - \mathbf{R}\right) = \mathrm{tr}\boldsymbol{\rho} - \mathrm{tr}\boldsymbol{\sigma} = 0$. This shows that $T = \mathrm{tr}Q = \mathrm{tr}P^+Q = \mathrm{tr}P^+(Q - R) =$

[41] More generally, the p-norm on operators is $\| Z \|_p \equiv \left(\mathrm{tr}\left(Z^\dagger Z\right)^{p/2}\right)^{1/p}$ and various p have various purposes.

[42] More explicitly: $\mathbf{Q}$ is the projector onto the subspace where $\boldsymbol{\rho} - \boldsymbol{\sigma}$ is positive. $\boldsymbol{\rho} - \boldsymbol{\sigma} = \mathbf{U}d\mathbf{U}^\dagger$ is hermitian and has a spectral decomposition; $\mathbf{Q} = \mathbf{U}d_+\mathbf{U}^\dagger$ is the bit with just the positive eigenvalues. So

$$\begin{aligned}
\boldsymbol{\rho} - \boldsymbol{\sigma} &= \mathbf{U}\ \mathrm{diag}(|d_1|, |d_2|...|d_n|, \ \ -|d_{n+1}|, \cdots \ \ -|d_d|)\ \mathbf{U}^\dagger, \\
\mathbf{Q} &= \mathbf{U}\ \mathrm{diag}(|d_1|, |d_2|...|d_n|, \ \ 0\ , \qquad \cdots \quad 0)\ \mathbf{U}^\dagger, \\
P^+ &= \mathbf{U}\ \mathrm{diag}(\ 1\ \ ,\ 1\ \ ...\ 1\ ,\ \ \ 0\ , \qquad \cdots \quad 0)\ \mathbf{U}^\dagger,
\end{aligned}$$

P^+ is the projector which will come up in all the calculations below. These manipulations are named after Hahn and Jordan.

$\mathrm{tr}P^+(\rho-\sigma) \geq \mathrm{tr}P(\rho-\sigma)$ for any other projector P, *i.e.*

$$T(\boldsymbol{\rho},\boldsymbol{\sigma}) = \max_P \mathrm{tr}P\left(\boldsymbol{\rho}-\boldsymbol{\sigma}\right) \tag{6.7}$$

where P is a projector, since the maximum is obtained when $P = P^+$ projects onto the same subspace as $\mathbf{Q}$. This is useful because it implies the triangle inequality for trace distance: take the $P = P^+$ which is the maximizer in (6.7), then add and subtract

$$T(\boldsymbol{\rho},\boldsymbol{\sigma}) = \mathrm{tr}P^+(\boldsymbol{\rho}-\boldsymbol{\sigma}) = \mathrm{tr}P^+(\boldsymbol{\rho}-\boldsymbol{\tau}) + \mathrm{tr}P^+(\boldsymbol{\tau}-\boldsymbol{\sigma}) \leq T(\boldsymbol{\rho},\boldsymbol{\tau}) + T(\boldsymbol{\tau},\boldsymbol{\sigma}).$$

A result which follows by the same logic is

$$T(\boldsymbol{\rho},\boldsymbol{\sigma}) = \max_{\{\mathsf{E}_x\}} T(p_x, q_x) \tag{6.8}$$

where $\{\mathsf{E}_x\}$ is a POVM and $p_x = \mathrm{tr}\mathsf{E}_x\boldsymbol{\rho}, q_x = \mathrm{tr}\mathsf{E}_x\boldsymbol{\sigma}$ are the resulting classical distributions, so that

$$T(p_x, q_x) \equiv \frac{1}{2}\sum_x |p_x - q_x| = \frac{1}{2}\sum_x \left|\mathrm{tr}\left(\mathsf{E}_x\left(\boldsymbol{\rho}-\boldsymbol{\sigma}\right)\right)\right| \tag{6.9}$$

is the classical trace distance. (Proof: The maximum is again obtained by including in the POVM a projector onto the support of $\mathbf{Q}$, whatever else is in $\{\mathsf{E}_x\}$ doesn't matter, so we may as well just take $\mathsf{E}_0 = P, \mathsf{E}_1 = \mathbb{1} - P$.) This says that two density matrices which are close together in trace distance give similar probability distributions for measurement outcomes.

Further, it gives an operational interpretation of the trace distance in terms of the optimal measurement to do if you must try to distinguish two states with a single measurement. More specifically, suppose at a random taste test you are given (with equal probability) one of two states, either $\boldsymbol{\rho}$ or $\boldsymbol{\sigma}$ and asked to guess which, and are allowed to perform only a single measurement. WLOG take the measurement E to be a two-outcome projective measurement: say 0 means you should guess $\boldsymbol{\rho}$ and 1 means you should guess $\boldsymbol{\sigma}$, *i.e.* $\mathrm{tr}\mathsf{E}_0\boldsymbol{\rho} \geq \mathrm{tr}\mathsf{E}_0\boldsymbol{\sigma}$ and $\mathrm{tr}\mathsf{E}_1\boldsymbol{\rho} \leq \mathrm{tr}\mathsf{E}_1\boldsymbol{\sigma}$. Then the probability of guessing right is

$$p_{\checkmark} = \frac{1}{2}\mathrm{tr}\mathsf{E}_0\boldsymbol{\rho} + \frac{1}{2}\mathrm{tr}\mathsf{E}_1\boldsymbol{\sigma} = \frac{1}{2}\left(1 + T(\mathsf{E}(\boldsymbol{\rho}),\mathsf{E}(\boldsymbol{\sigma}))\right) \overset{(6.8)}{\leq} \frac{1}{2}\left(1 + T(\boldsymbol{\rho},\boldsymbol{\sigma})\right).$$

In the second step we rewrote $\mathsf{E}_0 = \frac{1}{2}\left(\mathsf{E}_0 + \mathbb{1} - \mathsf{E}_1\right), \mathsf{E}_1 = \frac{1}{2}\left(\mathsf{E}_1 + \mathbb{1} - \mathsf{E}_0\right)$ and used (6.9) and the fact that $\mathrm{tr}\mathsf{E}_0\boldsymbol{\rho} \geq \mathrm{tr}\mathsf{E}_0\boldsymbol{\sigma}$ (and the reverse for 1).[43]

[43] Incidentally, the relative entropy also has an operational interpretation in terms of distinguishing states. Recall that classically $D(p|q)$ was how much longer our symbol code for p would be if we thought the distribution was q. This is the same as saying that the probability of mistaking p for q after sampling n trials is $e^{-nD(p|q)}$ (this is called Sanov's theorem by C&T). A similar statement is true quantumly:

$$\text{Prob(mistaking } \boldsymbol{\rho} \text{ for } \boldsymbol{\sigma} \text{ in } n \text{ measurements)} \simeq e^{-n\hat{D}(\boldsymbol{\rho}||\boldsymbol{\sigma})}.$$

Monotonicity of the trace distance. Now you will recall that we had to do some heavy lifting to show that the relative entropy was monotonic under quantum channels. For the trace distance, this is elementary:

$$T(\mathcal{E}(\boldsymbol{\rho}), \mathcal{E}(\boldsymbol{\sigma})) \leq T(\boldsymbol{\rho}, \boldsymbol{\sigma}). \tag{6.10}$$

Proof #1 of (6.10) (C&N p.407): In the notation of the previous calculations, trace-preserving means that $\text{tr}\mathcal{E}(\mathbf{Q}) = \text{tr}\mathbf{Q} = \text{tr}\mathbf{R} = \text{tr}\mathcal{E}(\mathbf{R})$. So

$$T(\boldsymbol{\rho}, \boldsymbol{\sigma}) = \text{tr}\mathbf{Q} = \text{tr}\mathcal{E}\left(\mathbf{Q}\right).$$

Now let P^+ be the projector which picks out the positive part of $\mathcal{E}(\boldsymbol{\rho} - \boldsymbol{\sigma})$, so

$$T(\mathcal{E}(\boldsymbol{\rho}), \mathcal{E}(\boldsymbol{\sigma})) = \text{tr}P^+\left(\mathcal{E}(\boldsymbol{\rho}) - \mathcal{E}(\boldsymbol{\sigma})\right) \leq \text{tr}P^+\mathcal{E}(\mathbf{Q}) \leq \text{tr}\mathcal{E}(\mathbf{Q}) = \text{tr}\mathbf{Q} = T(\boldsymbol{\rho}, \boldsymbol{\sigma}).$$

The two inequality steps use respectively the positivity of $\mathcal{E}(\mathbf{Q})$ (to say $\text{tr}P^+\mathcal{E}(\boldsymbol{\rho} - \boldsymbol{\sigma}) = \text{tr}P^+\mathcal{E}(\mathbf{Q} - \mathbf{R}) \leq \text{tr}P^+\mathcal{E}(\mathbf{Q})$) and of $\mathcal{E}(\mathbf{R})$, which in turn rely on the positivity of the channel $\mathcal{E}$. [End of Lecture 16]

Proof #2 of (6.10) (Christiandl §10): Write the Krauss representation of the channel: $\mathcal{E}(\boldsymbol{\rho}) = \sum_x \mathcal{K}_x \boldsymbol{\rho} \mathcal{K}_x^\dagger$. Then

$$\begin{aligned} T(\mathcal{E}(\boldsymbol{\rho}), \mathcal{E}(\boldsymbol{\sigma})) &= \frac{1}{2} \| \sum_x \left(\mathcal{K}_x \boldsymbol{\rho} \mathcal{K}_x^\dagger - \mathcal{K}_x \boldsymbol{\sigma} \mathcal{K}_x^\dagger\right) \|_1 \overset{\Delta}{\leq} \sum_x \frac{1}{2} \| \left(\mathcal{K}_x \boldsymbol{\rho} \mathcal{K}_x^\dagger - \mathcal{K}_x \boldsymbol{\sigma} \mathcal{K}_x^\dagger\right) \|_1 \\ &\overset{\text{c of t}}{=} \sum_x \frac{1}{2} \| \mathsf{E}_x \left(\boldsymbol{\rho} - \boldsymbol{\sigma}\right) \|_1 \overset{(6.8)}{\leq} T(\boldsymbol{\rho}, \boldsymbol{\sigma}) \end{aligned} \tag{6.11}$$

where $\mathsf{E}_x \equiv \mathcal{K}_x^\dagger P_x \mathcal{K}_x \geq 0$ and P_x is the projector onto $\mathcal{K}_x \boldsymbol{\rho} \mathcal{K}_x^\dagger - \mathcal{K}_x \boldsymbol{\sigma} \mathcal{K}_x^\dagger \geq 0$. 'c of t' stands for 'cylicity of the trace' and 'Δ' stands for triangle inequality.

Proof #3 of (6.10) (Preskill Chapter 2): As with the relative entropy, it suffices to prove monotonicity under partial trace. $T(\rho, \sigma)$ is the optimal distance between distributions of measurement results for distinguishing between ρ, σ. But the optimal measurement for distinguishing between ρ_A, σ_A is also a possible measurement for ρ_{AB}, σ_{AB} (it just doesn't act on B).

Strong convexity of the trace distance.

$$\begin{aligned} T\left(\sum_i^n p_i \boldsymbol{\rho}_i, \sum_j^n q_j \boldsymbol{\sigma}_j\right) &= \sum_i p_i \text{tr}P\boldsymbol{\rho}_i - \sum_i q_i \text{tr}P\boldsymbol{\sigma}_i \\ = \sum_i p_i \text{tr}P(\boldsymbol{\rho}_i - \boldsymbol{\sigma}_i) + \sum_i (p_i - q_i)\text{tr}P\boldsymbol{\sigma}_i &\leq \sum_i p_i T(\boldsymbol{\rho}_i, \boldsymbol{\sigma}_i) + T(p_i, q_i). \end{aligned}$$

P is the projector onto the positive subspace of $\sum_i (p_i \boldsymbol{\rho}_i - q_i \boldsymbol{\sigma}_i)$, $T(p_i, q_i)$ is the classical trace distance, and the inequality uses the relation (6.7). This implies joint convexity just by setting $p_i = q_i$! (The argument of problem 8 of HW 7 also shows that monotonicity implies joint convexity.)

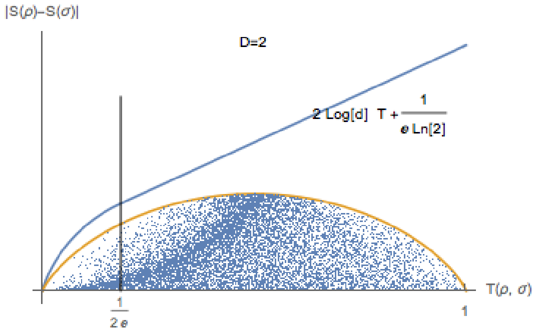
The exercises on page 408-409 of C&N make various interesting conclusions about the existence of fixed points of quantum channels from their ensmallening of the trace distance.

Fannes-Audenaert inequality: The von Neumann entropy is a continuous function on the space of density matrices because

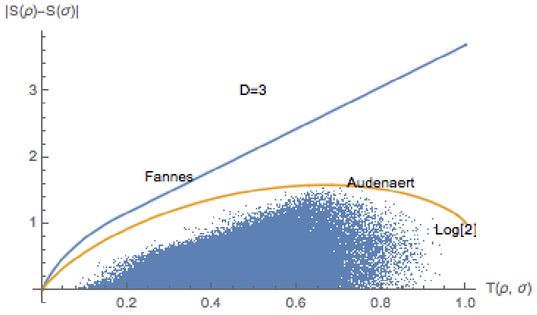
$$|S(\boldsymbol{\rho}) - S(\boldsymbol{\sigma})| \leq T(\boldsymbol{\rho}, \boldsymbol{\sigma}) \log(\mathfrak{D} - 1) + H_2(T(\boldsymbol{\rho}, \boldsymbol{\sigma})) \tag{6.12}$$

where $\mathfrak{D}$ is the dimension of the Hilbert space, and H_2 is the usual binary Shannon entropy function.

The dots in the figure are the entropy differences and trace distances of a collection of random density matrices (with dimension $n = \mathfrak{D} = 2$ here). The blue line in the figure at right is Fannes' bound, which while easier to prove (see C&N page 512), is visibly not tight. The yellow curve is Audenaert's improvement.



A notable feature of the yellow curve is that it goes down again when the trace distance is nearly maximal. Notice that $T(\boldsymbol{\rho}, \boldsymbol{\sigma}) \leq 1$ is saturated when the two states have orthogonal support. Having to leave room in $\mathcal{H}$ for the support of $\boldsymbol{\sigma}$ decreases the maximum entropy of $\boldsymbol{\rho}$. For the case of $\mathfrak{D} = 2$, two orthogonal states must both be pure. For $\mathfrak{D} > 2$, this is not the case, as you can see in the plot for $\mathfrak{D} = 3$ at right.



Both Fannes' and Audenaert's statements quickly reduce to classical statements about the eigenvalue vectors (p and q, respectively) of $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$: since $S(\boldsymbol{\rho})$ depends only on the eigenvalues, the LHS is $|S(\boldsymbol{\rho}) - S(\boldsymbol{\sigma})| = |H(p) - H(q)|$ and the only quantumness comes in the trace distance. But we saw already in (6.8) that the max defining the trace distance is realized by classical distributions. To be more precise, use the basis-independence to write

$$T(\boldsymbol{\rho}, \boldsymbol{\sigma}) = \frac{1}{2} \text{tr} |\Lambda_p - \mathbf{U} \Lambda_q \mathbf{U}^\dagger| \tag{6.13}$$

(where again Λ_p is the diagonal matrix with the eigenvalues p on the diagonal) and a result of Mirsky says

$$T(\text{eig}^\downarrow(A), \text{eig}^\downarrow(B)) \leq T(A, B) \leq T(\text{eig}^\downarrow(A), \text{eig}^\uparrow(B))$$

where $\text{eig}^{\downarrow}(A)$ means the list of eigenvalues of A in descending order. So the extremal values of (6.13) occur when $\mathbf{U}$ is a permutation matrix.

I'll go one more step: As usual in discussing trace distance, decompose $p - q \equiv q_+ - q_-$ where $q_\pm$ have support only when $p - q$ is $\gtrless 0$. Claim: The p, q which maximize $H(p) - H(q)$ at fixed $T(p, q)$ (a vertical line in the figure above) have $\text{rank}(q_+) = 1$, *i.e.* q_+ has only one nonzero entry, so that $T = \text{tr}q_+$. This is because $H(p) - H(q) = H(q + q_+ - q_-) - H(q)$ is concave in q_+ and the set of q_- (such that $\text{tr}q_+ = T, q_+ \geq 0, q_+q_- = q_-q_+ = 0$) is convex and therefore maxima must occur at the extremal points.

It seems like there should be a proof of the rest of the story from which one learns more but I haven't found it. However, the rest of the proof is actually constructive, and the result is that the inequality (6.12) is saturated for

$$\rho = \text{diag}(1, \underbrace{0, ..., 0}_{\mathfrak{D}-1}), \qquad \sigma = \text{diag}(1 - T, \underbrace{T/(\mathfrak{D} - 1), ...}_{\mathfrak{D}-1})$$

which have $T(\rho, \sigma) = \frac{1}{2}\| (-T, \underbrace{T/(\mathfrak{D} - 1), ...}_{\mathfrak{D}-1}) \|_1 = T$, and $S_1(\rho) = 0$ and $S_1(\sigma) = T \log(\mathfrak{D} - 1) + H_2(T)$.

Note that the analogous statement for the Renyi entropies with $\alpha > 1$ is *not* true: there are states which are close in trace distance with wildly different Renyi entropies. See appendix C of this monster for an illustration.

Trace distance bounds observable differences. If we know that two states are close in trace distance, we've seen that their entropies are also close. What about expectations of observables?[44] Indeed

$$|\langle \mathcal{O} \rangle_{\boldsymbol{\rho}} - \langle \mathcal{O} \rangle_{\boldsymbol{\sigma}}| \equiv |\text{tr}\,(\boldsymbol{\rho} - \boldsymbol{\sigma})\,\mathcal{O}| \overset{\triangle}{\leq} \underbrace{\text{tr}|\,(\boldsymbol{\rho} - \boldsymbol{\sigma})\,\mathcal{O}|}_{=\|\,(\boldsymbol{\rho}-\boldsymbol{\sigma})\mathcal{O}\,\|_1} \overset{\text{Hölder}}{\leq} \|\,\boldsymbol{\rho} - \boldsymbol{\sigma}\,\|_1 \|\,\mathcal{O}\,\| = 2\|\,\mathcal{O}\,\| T(\boldsymbol{\rho}, \boldsymbol{\sigma}).$$

The inequalities are the ordinary triangle inequality for the absolute value, and the Hölder inequality

$$\|\,XY\,\|_1 \leq \|\,X\,\|_p \|\,Y\,\|_q, \qquad p^{-1} + q^{-1} = 1$$

with $p = 1, q = \infty$ – note that $\|\,X\,\| = \|\,X\,\|_\infty$ is the largest eigenvalue of X when X is Hermitian.

A few words about the fidelity. [Christiandl, §10] What's bad about trace distance: it doesn't play well with purification and tensor products.

[44] Thanks to Wei-ting Kuo for asking about this.

If one or both of the states is pure, the fidelity $F(\boldsymbol{\rho},\boldsymbol{\sigma}) \equiv \| \sqrt{\boldsymbol{\rho}}\sqrt{\boldsymbol{\sigma}} \|_1$ (note that Preskill's $\sqrt{F}$ is my F!) reduces to more familiar (to me) things (since $\sqrt{|\psi\rangle\langle\psi|} = |\psi\rangle\langle\psi|$ for a 1d projector):

$$F(\boldsymbol{\rho},\boldsymbol{\sigma}) \overset{\text{if } \boldsymbol{\rho}=|\psi\rangle\langle\psi|}{=} \sqrt{\langle\psi|\,\boldsymbol{\sigma}\,|\psi\rangle} \overset{\text{if } \boldsymbol{\sigma}=|\phi\rangle\langle\phi|}{=} |\,\langle\psi|\phi\rangle\,|.$$

In fact, even for mixed states, the fidelity can be written like this:

$$F(\boldsymbol{\rho},\boldsymbol{\sigma}) = \max |\,\langle\sqrt{\boldsymbol{\rho}}|\sqrt{\boldsymbol{\sigma}}\rangle\,| \tag{6.14}$$

where the max is taken over purifications, $|\sqrt{\boldsymbol{\rho}}\rangle, |\sqrt{\boldsymbol{\sigma}}\rangle$, of the two states.

Here's why[45]: Let $|\Phi\rangle = \sum_k |kk\rangle$ be an (un-normalized) maximally entangled state, so

$$|\sqrt{\boldsymbol{\rho}}\rangle = \sqrt{\boldsymbol{\rho}}_A \otimes \mathbf{V}\,|\Phi\rangle\,, \quad |\sqrt{\boldsymbol{\sigma}}\rangle = \sqrt{\boldsymbol{\sigma}}_A \otimes \mathbf{W}\,|\Phi\rangle$$

are purifications of $\boldsymbol{\rho},\boldsymbol{\sigma}$, for any unitaries $\mathbf{V},\mathbf{W}$. Therefore:

$$\begin{aligned}\langle\sqrt{\boldsymbol{\rho}}|\sqrt{\boldsymbol{\sigma}}\rangle &= \langle\Phi|\,\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A} \otimes \mathbf{V}^\dagger\mathbf{W}\,|\Phi\rangle \\ &= \langle\Phi|\,\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A}\left(\mathbf{V}^\dagger\mathbf{W}\right)^t \otimes \mathbb{1}\,|\Phi\rangle \\ &= \mathrm{tr}\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A}\left(\mathbf{V}^\dagger\mathbf{W}\right)^t \overset{\text{polar}}{=} \mathrm{tr}|\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A}|\mathbf{U}\left(\mathbf{V}^\dagger\mathbf{W}\right)^t \le \mathrm{tr}|\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A}|\end{aligned}$$

where in the penultimate step we made a polar decomposition of $\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A} = |\sqrt{\boldsymbol{\rho}_A}\sqrt{\boldsymbol{\sigma}_A}|U$, and the last step is Cauchy-Schwartz inequality, with equality when $\mathbf{U}^\dagger = \left(\mathbf{V}^\dagger\mathbf{W}\right)^t$.

This result implies monotonicity of the fidelity under quantum channels, $F(\boldsymbol{\rho},\boldsymbol{\sigma}) \le F(\mathcal{E}(\boldsymbol{\rho}),\mathcal{E}(\boldsymbol{\sigma}))$, since the Stinespring dilation of $\mathcal{E}$ is one of the purifications over which we maximize in (6.14). Pretty slick.

I should probably mention that there are relations between fidelity and trace distance. First, for pure states, we saw $F(|\phi\rangle\langle\phi|, |\psi\rangle\langle\psi|) = |\,\langle\psi|\phi\rangle\,| \equiv |\sin\theta|$ (so that $\pi/2-\theta$ is the angle between the vectors – the problem can be reduced to the two dimensional plane spanned by ϕ,ψ). The trace distance in this case is

$$T(|\phi\rangle\langle\phi|, |\psi\rangle\langle\psi|)) = \frac{1}{2}\|\,|\phi\rangle\langle\phi| - |\psi\rangle\langle\psi|\,\|_1 = |\cos\theta|.$$

Therefore

$$T = \sqrt{1-F^2}, \quad \text{for pure states.} \tag{6.15}$$

More generally

$$1 - F(\rho,\sigma) \le T(\rho,\sigma) \le \sqrt{1-F(\rho,\sigma)^2}.$$

We can prove the upper bound by purifying both states:

$$F(\rho_A,\sigma_B)^2 = F(\rho_{AB},\sigma_{AB})^2 \overset{(6.15)}{=} 1 - T(\rho_{AB},\sigma_{AB})^2 \le 1 - T(\rho_A,\sigma_A)^2,$$

where the last step is monotonicity of the trace distance.

[45] This result is due to Uhlmann, and is what Preskill calls Uhlmann's Theorem.

6.5 Zoo of measures of mixed-state entanglement

Given a $\boldsymbol{\rho}_{AB}$ how do we check whether it is of the form (6.1)? One way to check is using a positive operator T on A which is not completely positive. Any such operator gives us $T \otimes \mathbb{1}_B$ which is positive on $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$, and it's positive on a convex combination of such states, hence on (6.1).

An example of a positive but not CP map is the partial transpose. On any operator on A, $X = \sum_{ij} X_{ij} |i\rangle\langle j|$ (written in some basis of A), T_A acts by $T_A(X) = \sum_{ij} |j\rangle\langle i|$. Partial transpose is the map which acts by $T_A \otimes \mathbb{1}_B$ on AB. More explicitly, choosing a basis of A and B, a separable state looks like

$$\rho_s = \sum_k p_k |e_k, f_k\rangle\langle e_k, f_k|$$

and its partial transpose is (using the fact that $X^T = X^{\star\dagger}$)

$$\rho_s^{T_A} = \sum_k p_k |e_k^\star, f_k\rangle\langle e_k^\star, f_k| \geq 0.$$

So the transpose operation in some basis of A is useful. Beware that there are examples of entangled states which are not identified as such by the transpose operation. In general, the CJ isomorphism maps positive but not completely positive operators to states called, naturally, 'entanglement witnesses'. For more on this see this review.

More ambitiously, we'd like to quantify how far from such a state is a given bipartite state $\boldsymbol{\rho}_{AB}$. Some desiderata for such an entanglement measure $E(\boldsymbol{\rho}_{AB})$ are [Vedral, quant-ph/0102094]:

1. The measure $E \geq 0$ vanishes for separable states: $E(\sum_a p_a \rho_a^A \otimes \rho_a^B) = 0$. We might further like it if $E = 0$ *only* for such states.

2. E is invariant under local basis changes: $E(\rho) = E(U_A \otimes U_B \rho U_A^\dagger \otimes U_B^\dagger)$.

3. E doesn't increase under LOCC. A slightly stronger but easier to study condition is that E doesn't increase under 'locally separable operations,' which just means a channel where the Kraus operators factorize:

$$E(\rho) \geq \sum_a p_a E(A_a \otimes B_a \rho A_a^\dagger \otimes B_a^\dagger / p_a)$$

   where A, B are local operators on A, B. (Actually, 3 implies 2 since if E decreases under U, it must increase under $U^{-1} = U^\dagger$.)

4. If ρ_{AB} happens to be pure, it would be nice if E coincides with the entanglement entropy: $E(|\psi\rangle\langle\psi|) = S(\rho_A)$.

Some solutions to these conditions are the following (there are many, so I guess we need more conditions):

Entanglement of distillation. We've already talked about this as an entanglement measure:

$$E_D(\rho) \equiv \lim_{n\to\infty} \frac{m}{n}$$

where m is the maximum number of singlets (maximally entangled qbit pairs) which can be distilled from n copies of ρ, as we did above for pure states in §6.3. In §6.3 we showed (by combining Schumacher compression and quantum teleportation) that for pure states $E_D(\rho) = S_A$, *i.e.* condition 4.

Entanglement of formation. Consider all decompositions of

$$\rho_{AB} = \sum_a p_a |\psi_a\rangle\langle\psi_a| \tag{6.16}$$

into pure states and minimize the average entropy:

$$E_F(\rho) \equiv \min_{p,\psi} \sum_a p_a S(\rho_A^a)$$

where $\rho_A^a = \mathrm{tr}_B |\psi_a\rangle\langle\psi_a|$.

Claim: This has an operational interpretation in terms of the reverse process of distillation: $E_F(\rho^{\otimes n})$ is asymptotically the minimum number of Bell pairs required to make n copies of ρ by LOCC. (C&N call this process 'entanglement dilution'. And this reverse measure is called the *entanglement cost*, whose full definition is this horror:

$$E_C(\rho) \equiv \inf\{r : \lim_{n\to\infty} \inf_{\mathcal{L}} T(\rho^{\otimes n}, \mathcal{L}(\Phi(2^{rn}))) = 0\}$$

where $\mathcal{L}$ is a LOCC operation, $\Phi(2^m)$ is a state of m Bell pairs, and T is a distance measure between states. It is sometimes useful to relax the demand that the distance be exactly zero.) For pure states, this number is $S(A)$ again, just like for distillation, basically by running the argument backwards (because the Shannon bound is tight on both sides). For mixed states, distillation and formation are not quite the same because the process is not reversible.

Here's why E_F is related to this process: suppose you have a convex decomposition of $\boldsymbol{\rho}$ into pure states, like (6.16). Then a protocol for making $\boldsymbol{\rho}$ from singlets is: sample the distribution to choose a particular pure state $|\psi_a\rangle$ with probability p_a. Then do the pure-state entanglement dilution protocol – making $|\psi_a\rangle$ requires $S_A(\rho_a)$ singlets. Then the average number of singlets required is $\sum_a p_a S_A(\rho_a)$. The smallest possible is obtained by minimizing over p_a, ψ_a. A rigorous argument that this average value is the asymptotic value is given in quant-ph/0008134.

Relative entropy of entanglement. This is directly measuring the 'distance' to the nearest disentangled state:

$$E_R(\rho) \equiv \min_{\sigma \in D} D(\rho||\sigma)$$

where D is the (convex) set of separable states.

Claim without proof: the three measures just introduced satisfy

$$E_D(\rho) \leq E_R(\rho) \leq E_F(\rho).$$

Squashed entanglement. [Preskill chapter 10]

$$E_{sq}(\rho_{AB}) \equiv \frac{1}{2} \inf_{\rho_{ABC}} \{I(A:B|C)|\rho_{AB} = \mathrm{tr}_C \rho_{ABC}\} \overset{\text{SSA}}{\geq} 0.$$

The infimum over extensions of ρ_{AB} squashes out the non-quantum correlations. If ρ_{AB} is pure, then any extension is $|\psi\rangle\langle\psi| \otimes \rho_C$, and $E_{sq}(|\psi\rangle\langle\psi|) = \frac{1}{2} I(A:B) = S_A$. If $\rho_{AB} = \sum_c p_c \rho_A^c \otimes \rho_B^c$ is separable, then the extension

$$\rho_{ABC} = \sum_c p_c \rho_A^c \otimes \rho_B^c \otimes |c\rangle\langle c|$$

(with $|c\rangle$ orthonormal) has $I(A:B|C) = 0$ (since $I(A:B|c) = 0$ for each value of c, and $I(A:B|C)$ is their average against p_c).

Claim: E_{sq} is monotonic under LOCC.

Proof [from the paper by Christiandl and Winter that introduced E_{sq}]: We'll show that $E_{sq}(\rho_{AB}) \geq \sum_k p_k E_{sq}(\tilde{\rho}_{AB}^k)$ where $\tilde{\rho}^k$ is the state resulting from a measurement by A which obtains outcome k. This result plus convexity of E_{sq} implies monotonicity under LOCC.

The idea is to treat A's measurement by dilation: introduce ancillary spaces A_0, A_1 in reference state $|00\rangle\langle 00|$, perform a unitary $U_{A_0A_1A}$ (note that it does not act on B), and trace out A_1. The output of the measurement is recorded in A_0:

$$\rho_{AB} \overset{\text{dilate}}{\mapsto} \rho_{AB} \otimes |00\rangle\langle 00|_{A_0A_1} \overset{\text{unitary}}{\mapsto} U_{A_0A_1A} \rho_{AB} \otimes |00\rangle\langle 00|_{A_0A_1} U^\dagger_{A_0A_1A} \overset{\mathrm{tr}_{A_1}}{\mapsto} \tilde{\rho}_{A_0AB} = \sum_k p_k |k\rangle\langle k|_{A_0} \otimes \tilde{\rho}_{AB}^k.$$

For any extension ρ_{ABC}, the conditional mutual information is invariant under extension by product states and local unitaries:

$$I_\rho(A:B|C) = I_{|00\rangle\langle 00|\otimes\rho}(AA_0A_1 : B|C) = I_{U_{A_0A_1A}|00\rangle\langle 00|\otimes\rho U^\dagger_{A_0A_1A}}(AA_0A_1 : B|C) \tag{6.17}$$

$$\overset{\text{MRE}}{\geq} I_{\tilde\rho}(AA_0 : B|C) \tag{6.18}$$

$$\overset{\text{chain rule}}{=} \underbrace{I_{\tilde\rho}(A_0 : B|C)}_{\geq 0} + I_{\tilde\rho}(A : B|A_0C) \tag{6.19}$$

$$\geq I_{\tilde\rho}(A : B|A_0C) = \sum_k p_k I_{\tilde\rho^k}(A : B|C) \geq 2\sum_k p_k E_{sq}(\tilde\rho_k). \tag{6.20}$$

A comment about the step labelled 'MRE': unlike the classical case, the quantum conditional mutual information is not itself a relative entropy. Rather, $I(A:B|C) = I(A:CB) - I(A:B)$ is a *difference* of relative entropies. But it is true that SSA implies

$$I(AA' : B|C) \geq I(A : B|C)$$

i.e. that the quantum CMI decreases upon discarding some degrees of freedom. Explicitly:

$$\begin{aligned} I(AA' : B|C) - I(A : B|C) &= I(AA' : BC) - I(AA' : B) - (I(A : BC) - I(A : B)) \\ &= S(AA') + S(BC) - S(AA'BC) - (S(AA') + S(B) - S(AA'B)) \\ &\quad - (S(A) + S(BC) - S(ABC) - (S(A) + S(B) - S(AB))) \\ &= S(ABA') + S(ABC) - S(AB) - S(ABA'C) \geq 0 \end{aligned} \tag{6.21}$$

where the last step is SSA.

To see convexity of E_{sq}, let ρ^a_{ABC} be extensions of ρ^a_{AB} (WLOG) by the same C. Then $\tau_{AB} \equiv \sum_a p_a \rho^a_{AB}$ is extended by

$$\tau_{ABCE} \equiv \sum_a p_a \rho^a_{ABC} \otimes |a\rangle\langle a|_E.$$

But

$$\sum_a p_a I_{\rho^a}(A : B|C) = I_\tau(A : B|CE) \geq 2E_{sq}(\tau_{AB})$$

which implies

$$\sum_a p_a E_{sq}(\rho^a) \geq E_{sq}\left(\sum_a p_a \rho^a\right).$$

One nice use of the squashed entanglement is to show the property of *monogamy of entanglement*. Classical correlations are polygamous in the sense that arbitrarily many parties can be strongly correlated with each other, for example by subscribing to the same twitter feed. No cloning prevents the analogous copying of quantum correlations. There is a reason a Bell pair is called 'maximally entangled'. Here is a quantification of this statement:

$$E_{sq}(A:B) + E_{sq}(A:C) \leq E_{sq}(A:BC) \leq \log \dim A. \tag{6.22}$$

[End of Lecture 17]

Proof: Recall the chain rule for mutual information:

$$I(A:BC) = I(A:C) + I(A:B|C)$$

and

$$I(A:BC|D) = I(A:C|D) + I(A:B|CD). \tag{6.23}$$

$$E_{sq}(A:BC) \equiv \frac{1}{2}\inf\{I(A:BC|D), \rho_{ABC} = \text{tr}_D \rho_{ABCD}\}$$

but any ρ_{ABCD} participating in the infimum is also an extension of $\rho_{AB} = \text{tr}_{CD}\rho_{ABCD}$ and $\rho_{AC} = \text{tr}_{BD}\rho_{ABCD}$, and therefore (6.23) (times $\frac{1}{2}$) says (6.22). The inequality is saturated by a pure state of the form $|\psi_{ABC}\rangle = |\psi_{A_L B}\rangle \otimes |\psi_{A_R C}\rangle$. ■

A few more properties of E_{sq} merit mention. As you'll show on the homework, E_{sq} is additive in the sense that $E_{sq}(\rho_{AB} \otimes \rho_{CD}) = E_{sq}(\rho_{AB}) + E_{sq}(\rho_{CD})$ (more generally it is superadditive $E_{sq}(\rho_{ABCD}) \geq E_{sq}(\rho_{AB}) + E_{sq}(\rho_{CD})$). These features imply

$$E_D \leq E_{sq} \leq E_C. \tag{6.24}$$

Combining with (6.22) then implies a monogamy relation for E_D and E_C: $E_D(A:B) + E_D(A:C) \leq E_C(A:BC)$.

Proof of (6.24). First let's see that $E_{sq}(\rho) \leq E_F$ (and hence E_C). Let $\rho_{AB} = \sum_k p_k |\psi_k\rangle\langle\psi_k|$ be the pure-state decomposition of ρ which minimizes $\inf \sum_k p_k S_{\psi_k}(A) = E_F(\rho)$.[46] Then consider (again) the extension

$$\rho_{ABC} = \sum_k p_k |\psi_k\rangle\langle\psi_k| \otimes |k\rangle\langle k|_C, \quad \{|k\rangle_C\}\text{are ON}.$$

Then

$$E_{sq}(\rho) \leq \frac{1}{2} I(A:B|C) = \frac{1}{2}\sum_k p_k I_{\psi_k}(A:B) \overset{\text{purity}}{=} \sum_k p_k S_{\psi_k}(A) \leq E_F(\rho).$$

[46] You could give me a hard time about the fact that the infimum may not be realized by any actual set of states. And I would say: what we'll show is that E_F is the infimum over a certain set of extensions of ρ_{AB}, while E_{sq} is the infimum over all extensions, so E_{sq} is smaller.

This implies that $E_{sq} \geq E_C$ because

$$E_C(\rho) = \lim_{n\to\infty} E_F(\rho^{\otimes n})/n \geq E_{sq}(\rho^{\otimes n})/n = E_{sq}(\rho)$$

by the additivity property of E_{sq} for product states.

The lower bound $E_{sq} \geq E_D$ follows from monotonicity under LOCC. Suppose the most efficient distillation process takes $\rho_{AB}^{\otimes n}$ to $|s\rangle\langle s|$ where $|s\rangle$ is a maximally entangled state of rank s. Then $E_D(\rho) = \log(s)/n$. Then

$$nE_{sq}(\rho) \overset{\text{additivity}}{=} E_{sq}\left(\rho^{\otimes n}\right) \overset{\text{LOCC monotone}}{\geq} E_{sq}(|s\rangle\langle s|) = \log s = nE_D(\rho).$$

[47] This part is true for any continuous additive LOCC monotone.

Entanglement of purification. [For a recent discussion of its application to quantum many body physics, I recommend Nguyen et al]

$$E_P(\rho) \equiv \min_{\psi, A'} S_{AA'}$$

where $|\psi\rangle \in AA'BB'$ is a purification of ρ.

This actually fails condition 1 completely:

$$E_P\left(\sum_a p_a|a\rangle\langle a| \otimes |a\rangle\langle a|\right) = H(p).$$

E_P also has an operational interpretation in terms of local operations and communication on asymptotically many copies of the state. It satisfies:

$$I(A:B)/2 \leq E_P(A:B) \leq \min(S_A, S_B)$$

$$E_P(A:BC) \geq E_P(A:B)$$

In practice, the previous measures of entanglement all have the shortcoming that their computation requires some minimization over some enormous space. The next example has the advantage of being somewhat more practical (though still difficult) to compute.

Entanglement negativity. This is a direct attempt [Vidal-Warner] to quantify the fact that the partial transpose is an entanglement witness. Most directly we can

[47] Again the brevity of this proof is at the cost of some robustness: One might want to allow the LOCC process to take $\rho^{\otimes n}$ to some σ which is merely *close* to $|s\rangle\langle s|$. In that case the result relies on the *continuity* of E_{sq} (two states which are close have similar E_{sq}). (There is a concise proof of continuity of E_{sq} in Petz' book, page 69.) Also, again the inf may not be realized by any actual $|s\rangle$.

just sum the negative eigenvalues of ρ^{T_A} where $\rho^{T_A} \equiv (T_A \otimes 1\!\!1_B)(\rho)$ is the partial transpose:

$$\mathcal{N}(\rho) \equiv \frac{1}{2}\left(\| \rho^{T_A} \|_1 - 1\right).$$

$\mathcal{N}(\rho)$ is the sum of the negative eigenvalues of ρ^{T_A}. To see this, note (again, as in §6.4) that any hermitian operator A (such as ρ^{T_A}) can be decomposed as a difference of positive operators

$$A = a_+\rho^+ - a_-\rho^- \tag{6.25}$$

where $\rho^\pm$ are density matrices. Then $\text{tr}A = a_+ - a_-$. There is a *best* such decomposition, which is when the support of $\rho^\pm$ are the eigenspaces of A with positive and negative eigenvalues (as in §6.4). In that case, $\| A \|_1 = a_+ + a_-$, and $a_- = -\text{tr}P^-A$, where P^- is the projector onto the negative eigenspaces of A. $A = \rho^{T_A}$ still has unit trace, so $1 = a_+ - a_-$ and we find $\| A \|_1 = 1 + 2a_-$. Hence

$$a_- = -\text{tr}\left(AP^-\right) = \mathcal{N}(A) = \sum |\text{negative eigenvalues of } A|.$$

More generally, the decomposition (6.25) means $0 \leq A + a_-\rho^-$ which implies

$$0 \leq \text{tr}\left((A + a_-\rho^-)P^-\right) = -\mathcal{N} + a_- \underbrace{\text{tr}\left(P^-\rho^-\right)}_{\leq 1} \quad \Longrightarrow \; a_- \geq \mathcal{N}.$$

The bound is saturated when $a_-\rho^- = -P^-AP^-$, that is, when ρ^- is the negative-eigenvalue part of A. Therefore

$$\mathcal{N}(\rho) = \inf_{a_\pm,\rho_\pm}\{a_- | \rho^{T_A} = a_+\rho_+ - a_-\rho_-\}.$$

$\mathcal{N}$ is convex just because of the triangle inequality on the trace norm.

The proof that $\mathcal{N}$ is monotonic under LOCC is nice. It suffices to consider a single element of the family $\{\mathcal{M}_a\}$ of CP maps describing the local measurement and CC: $\rho \to \sum_a \mathcal{M}_a(\rho) = \sum_a p_a\rho_a'$. Local measurement means $\mathcal{M}_a(\rho) = (1\!\!1_A \otimes M_a)\,\rho\left(1\!\!1_A \otimes M_a^\dagger\right)$. This implies that $\mathcal{M}$ commutes with partial transpose: $\mathcal{M}_a(\rho)^{T_A} = \mathcal{M}_a(\rho^{T_A})$. Then given the optimal decomposition $\rho^{T_A} = (1+N)\rho^+ - N\rho^-$ (where N is the initial negativity), the partial transpose of the output is

$$p_a\left(\rho_a'\right)^{T_A} = \mathcal{M}_a(\rho^{T_A}) = (1+N)\mathcal{M}_a(\rho^+) - N\mathcal{M}_a(\rho^-)$$

which (after dividing by p_a) is a decomposition of the form (6.25). Hence $\mathcal{N}(\mathcal{M}_a(\rho)) \leq N/p_a$. Therefore

$$\sum_a p_a\mathcal{N}(\rho_a') \leq \mathcal{N}(\rho).$$

A related quantity which is additive under composition ($E_N(\rho_1 \otimes \rho_2) = E_N(\rho_1) + E_N(\rho_2)$) is the *logarithmic negativity*

$$E_N(\rho) \equiv \log \| \rho^{T_A} \|_1 = \log \| \rho^{T_B} \|_1 = \log \sum_k |\lambda_k|$$

The λ_k are the eigenvalues of ρ^{T_A}. The logarithmic negativity satisfies conditions 1-3, but not 4 (for a pure state, it is the Renyi entropy of index 1/2 rather than the vN entropy). It is also not convex or concave. Being an LOCC monotone, it also bounds $E_D(\rho)$ from above.

But a virtue of the negativity is that it can be seen to throw away classical correlations. In particular, it does not see thermal entropy. To understand this consider the thermal density matrix of a many body system

$$\rho = e^{-\beta H}/Z = e^{-\beta(H_A + H_B + H_{AB})}/Z$$

where the subscripts on the terms in H indicate their support. If we throw away the boundary terms H_{AB}, this would be $\rho_0 = e^{-\beta H_A} e^{-\beta H_B}/Z_0$, which has zero negativity. This suggests that (the logarithmic) negativity should satisfy an area law for local Hamiltonians where H_{AB} contains only an area law's worth of terms. This statement is proved in Appendix A of this paper.

In summary, here is a table from 1010.1750 (whose purpose is to show that squashed entanglement vanishes *only* for separable states, which is what is meant by *faithfulness*):

Measure	E_{sq} [1]	E_D [2, 3]	K_D [4, 5]	E_C [6, 2]	E_F [2]	E_R [7]	E_R^∞ [8]	E_N[9]
normalisation	y	y	y	y	y	y	y	y
faithfulness	y	n [10]	?	y [11]	y	y	y [12]	n
LOCC monotonicity	y	y	y	y	y	y	y	y [16]
asymptotic continuity	y [17]	?	?	?	y	y [18]	y [13]	n[13]
convexity	y	?	?	?	y	y	y [19]	n
strong superadditivity	y	y	y	?	n [20, 21]	n [22]	?	?
subadditivity	y	?	?	y	y	y	y	y
monogamy	y [15]	?	?	n [14]	n [14]	n [14]	n [14]	?
computability!	n	n	n	n	n	n	n	y

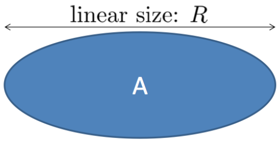
Table 2: [adapted from 1010.1750]. If no citation is given, the property either follows directly from the definition or was derived by the authors of the main reference. Many recent results listed in this table have significance beyond the study of entanglement measures, such as Hastings's counterexample to the additivity conjecture of the minimum output entropy [21] which implies that entanglement of formation is not strongly superadditive [20].

7 Area laws and local tensor network states

Now we incorporate some notion of spatial locality into our quantum systems: imagine that the Hilbert space is a tensor product over patches of d-dimensional space, and imagine that the system is governed by a local Hamiltonian $H = \sum_x H_x$. Our job now is to understand the consequences of locality for the physics of such a system.

Expectations. We'll begin with some *facts*, not all of which have been proved by humans so far. Then we will come back more systematically and see which of them we can understand with our brains and the tools we've been developing.

Everywhere in this discussion we will talk about a subregion of linear size R (think of R as the diameter), and we will be interested in the scaling with R. So Volume$(A) \sim R^d$.



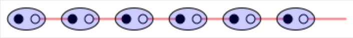
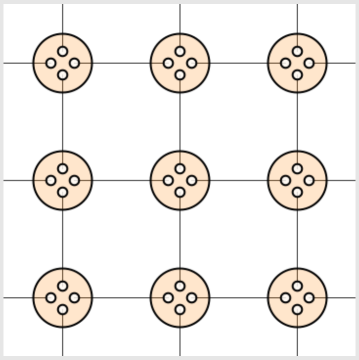
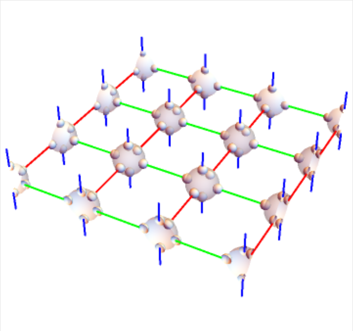
A basic expectation is that groundstates of local hamiltonians $H = \sum_x H_x$ have area law entanglement. In d spatial dimensions, this means that a subregion of linear size R will have an entanglement entropy whose largest term scales like R^{d-1}, when $R \gg a$, the lattice spacing.

Very roughly, the idea is that the minimizing the energy involves strongly entangling sites with participate in each term H_x, but we do not expect such large entanglement between distant sites. When we cut out a region, we cut entanglement bonds mostly between the sites in a thin layer near the boundary of the region, and their number scales like R^{d-1} (for $R \gg a$). This intuition also shows that the coefficient of the area-law term depends on UV details – the area law term is an obstacle to extracting possibly-universal terms subleading in the large-R expansion. More on this below.

Example with exact area law. The area law is motivated by the fact that if the whole system is in a pure state, entanglement arises only by cutting entanglement bonds between the subsystem and its complement, and in groundstates of local Hamiltonians, those bonds are mostly between nearest neighbors. Here is an example where this intuition is exactly true:

Consider the Heisenberg antiferromagnetic interaction between two spin $\frac{1}{2}$s: $H_{ij} = J\left(X_iX_j + Y_iY_j + Z_iZ_j\right)$. The groundstate is the spin singlet. The spin triplet has energy larger by J. Think of J as big. So the groundstate of this hamiltonian is a maximally entangled state between the two spins at the ends.

Now imagine that each site is made of a cluster of spin $\frac{1}{2}$s, one for each end of a link ending at that site. For hypercubic lattices in $d = 1, 2, 3$ this looks like this:

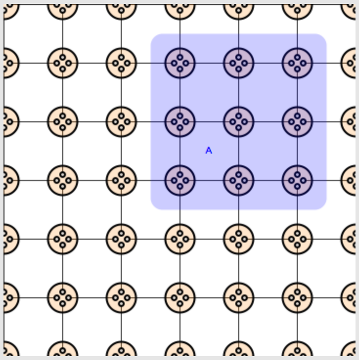




For example, for the square lattice we have four qbits per site. (These can be organized by their spin, and all the action is actually in the spin-2 subspace, but this is not essential to the point I am trying to make.) Now let $H = \sum_{\text{bonds}\langle ij\rangle} H_{ij}$. The groundstate, by design, is

$$|\text{gs}\rangle = \otimes_{\text{bonds}} \frac{|\uparrow_i\downarrow_j\rangle - |\downarrow_i\uparrow_j\rangle}{\sqrt{2}}.$$

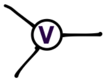
The terms in H all commute since they act on different spins. The first excited state is obtained by breaking any singlet. (If one wants to make a more physical model where things can move around, it is a good idea to add terms which penalize the spin-0 and spin-1 states of each site. Projecting onto the symmetric combination at each site (spin $z/2$ for coordination number z) results in the AKLT model.)

In this model, the entanglement entropy of a subregion is exactly equal to the number of bonds crossing its boundary.



7.1 Local tensor network states

The following may be regarded as a solution to the area law condition: again we draw feynman diagrams, and we associate wavefunctions to diagrams; each leg is associated with a Hilbert space; if we choose a basis, we get a (complex-number-valued) tensor. Dangling legs indicate free indices. So a tensor $V_{i_1\cdots i_k}$ is associated with a state in

$$\mathcal{H}^{\otimes k} \ni |V\rangle = \sum_{i_1\cdots i_k}^{\chi} V_{i_1\cdots i_k} |i_1 \cdots i_k\rangle =$$


($k = 3$ legs in the diagram).

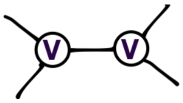
Previously, we've considered legs connecting tensors to be contracted by δ_{ij}, but it can be useful also to think of the links in the diagram as (maximally-entangled) states, by the usual isomorphism between End$\mathcal{H} \simeq \mathcal{H} \otimes \mathcal{H}^\star$:

$$\langle L| = \sum_{ij} \langle ij| \Phi_{ij}, \qquad e.g., \ \Phi_{ij} = \frac{\delta_{ij}}{\sqrt{\chi}}.$$

So a state associated with a graph Γ can be written as

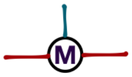
$$|\Gamma\rangle = \left(\otimes_{\text{links},L} \langle L|\right) \left(\otimes_{\text{vertices, } v} |T_v\rangle\right) \in \otimes_{\text{dangling legs}} \mathcal{H}.$$

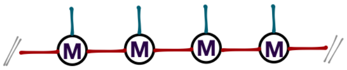
For example:



$$= \langle L| \left(|V\rangle \otimes |V\rangle\right) = \sum_i V_{ijk} V_{ilm} |jklm\rangle .$$

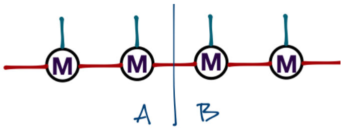
Generally this construction, with the link states, is called PEPS (projected entangled pair states). [See this recent paper for a clear recent account of this point of view.]

Sometimes it is useful to distinguish some of the indices, *i.e.* to regard one of the indices at each site as living in the actual space, and the rest as auxiliary, *e.g.*  where the blue index lives in $\mathcal{H}_x$ (goes up to $\mathfrak{D}$), while the red indices are auxiliary, $a, b = 1..\chi$, the *bond dimension.*



Then we can make a state in $\otimes_x \mathcal{H}_x$ by contracting the auxiliary indices (for example, with periodic boundary conditions in one dimension, as at right).

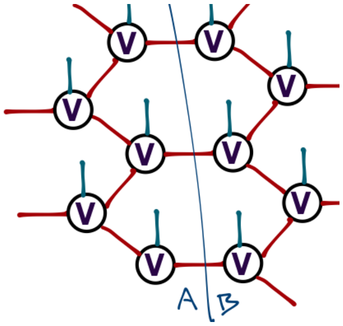
A state constructed in this way automatically satisfies the area law. Consider a left-right bipartition in the previous figure (consider open boundary conditions here). The state is basically already in the Schmidt representation, in the sense that it is of the form



$$\sum_{a_i=1}^{\chi} \left(...M^{\sigma_{i-1}}_{a_{i-1}a_i} \left|\sigma^1 \cdots \sigma^{i-1}\right\rangle \right) \otimes \left(M^{\sigma_i}_{a_i a_{i+1}} \cdots \left|\sigma^i \cdots \sigma^N\right\rangle \right)$$

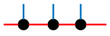
– a sum of χ rank one matrices. χ is then an upper bound on the Schmidt rank, and we can bound the entanglement entropy above by $\log \chi$.

More generally, in higher dimensions, we can bound the entropy of a subregion by $\log \chi$ times the number of bonds crossed by the entangling surface ($3 \log \chi$ for the figure at right).



Conversely, an exact area law *implies* such a local tensor network state: an area law means we can do Schmidt decomposition across every cut, and we need at most $\chi = \mathcal{O}(L^0)$ singular values for each bond, a number which does not grow with system size.

In 1d, such a local tensor network state is called a matrix product state (MPS):



$$= \sum_{a_{1,2...}=1}^{\chi} A^{\sigma_1}_{a_1a_2} A^{\sigma_2}_{a_2a_3} \cdots \left|\sigma_1, \sigma_2 \cdots\right\rangle \qquad \chi \equiv \text{bond dimension} \tag{7.1}$$

When $\chi = 1$, it is a product state. More generally χ is the Schmidt rank of the state for any bipartition (with open boundary conditions, so that we only have to cut once to cut the system in two parts).

AKLT. For example, if we take all the matrices to be the same, and take the minimal $\chi = 2$ and the local Hilbert space dimension to be 3 (*i.e.* a spin 1 at each site), and set

$$M^{0} = \boldsymbol{\sigma}^x/\sqrt{2}, \;\; M^{1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \;\; M^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

and take each of the link states to be a singlet $\langle L| = \langle ab| \mathbf{i}\sigma^y_{ab}$ (a particular maximally entangled state chosen for its $\mathsf{SU}(2)$ spin invariance), we get the AKLT state. (So the tensors A in (7.1) are $A = M\mathbf{i}\sigma^2$.) This is really just group theory: in the tensor product of two spin-$\frac{1}{2}$s ($2 \times 2 = 1 + 3$) the tensor which projects onto the triplet (the symmetric part) is

$$M^{\sigma}{}_{ab} |\sigma\rangle \langle ab| = \langle 0| \left(|\uparrow\uparrow\rangle + |\downarrow\downarrow\rangle\right)/\sqrt{2} + |1\rangle\langle\uparrow\uparrow| + |-1\rangle\langle\downarrow\downarrow|.$$

See this paper and this one for more examples of matrix product states.

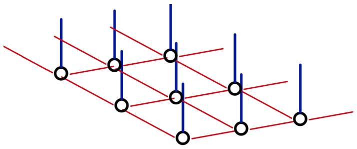
It's not necessary that all the tensors be the same; we could have labelled them $(M_i)^{\sigma_i}_{a_i a_{i+1}}$ so they depend on their site. Note that there is some ambiguity in the Ms comprising an MPS: if I change

$$(M_i)^{\sigma_i}_{a_i a_{i+1}} \mapsto U^{(i)}_{a_i a'_i} (M_i)^{\sigma_i}_{a'_i a'_{i+1}} U^{(i+1)\dagger}_{a'_{i+1} a_{i+1}}$$

then the state does not change (assuming we take care of the ends of the chain, *e.g.* by periodic boundary conditions). [End of Lecture 18]

Regarding (7.1) as a variational ansatz, and minimizing the energy expectation over the values of the matrices M gives a version of the DMRG ('density matrix renormalization group') algorithm, which is a very popular numerical method for interacting systems, and which works well in low dimensions. For more, see the review DMRG in the age of MPS. (See §4.1.5 of that paper for more detail on the AKLT state, too.)

In 2d, the solution of the area law looks something like the network (PEPS) at right. It is not so easy because in $d > 1$ even an area law grows with system size.



Exceptions to the area law. The 'expectations' above are often correct (even beyond the examples where they are precisely true), but there are some real exceptions to the area law expectation, even for groundstates of local Hamiltonians: groundstates at quantum critical points in $d = 1$ have

$$S_{1d \text{ critical}}(A) = \frac{c}{3} \log R/\epsilon, \tag{7.2}$$

whereas the $d = 1$ area law would be independent of R. ϵ is a short-distance cutoff, and c, the 'central charge', is a measure of the number of degrees of freedom. This includes the well-studied case of $1 + 1$-dimensional conformal field theory (CFT), where much can be said (if you are impatient, look here). A positive way to look at this is that the entanglement entropy of subregions can diagnose a continuous phase transition.

In $d > 1$, even critical points (are expected to) satisfy the area law. An important class of exceptions to the area law in any dimension is metallic groundstates of fermions, such as free fermions in partially-filled bands. This also leads to a logarithmic violation: $S_{\text{metal}}(R) \sim (k_F R)^{d-1} \log(k_F R)$ super-area-law scaling, where the Fermi momentum k_F makes up the dimensions. This result can be understood from the $1+1$-d CFT case, as explained here. Another class of examples of area law violation in $d = 1$ arises from highly disordered systems, namely random singlet states. More on this on the homework.

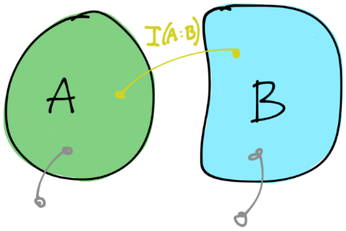
Non-groundstates. And of course there is more in the world than groundstates. The first excited state of a many body system has zero energy density in the thermodynamic limit. (Often it is a single particle.) Such states will still have an area law if the groundstate did. But in general, states with finite energy density and certainly finite temperature states violate the area law! Look again at a thermal double purification of some thermal state:

$$|\sqrt{\boldsymbol{\rho}}\rangle = Z^{-1/2} e^{-\frac{1}{2}\beta \mathbf{H} \otimes \mathbb{1}} \sum_i \frac{1}{\sqrt{|\mathcal{H}|}} |ii\rangle = Z^{-1/2} \sum_E e^{-\frac{1}{2}\beta E} |EE\rangle . \tag{7.3}$$

The maximally entangled state here can be in any basis; we can choose it to be a local basis: each site is strongly entangled with its purifying partner. The ancilllary Hilbert space doing the purifying is really just a proxy for the thermal bath and we are using our freedom to mess with the purification to make it look nice (like a copy of our system). (Beware that the individual object in the intermediate step I've written in (7.3) are maybe not so well-defined in the thermodynamic limit.) The lesson I am trying to convey is: Volume law entanglement entropy of thermal states can be regarded as entanglement with the thermal bath, rather than entanglement between parts of the system itself.

7.2 Mutual information appreciation subsection

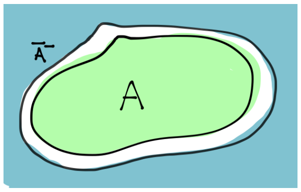
[Most of the discussion in this section follows Wolf-Verstraete-Hastings-Cirac ($\equiv$ WVHC)] We have seen above the utility of mutual information in Shannon theory. Mutual information also has many virtues in quantum many body physics.



• **Mutual information quantifies only correlations, no entropy of mixture.** A nice virtue arises when we think about mixed states of the full system: the mutual information between two subsystems subtracts out the entanglement with the environment. On the other hand, when the full state is pure, it reduces to $I(A:B) \overset{\text{AB pure}}{=} 2S(A)$.

• **Mutual information of separated regions is UV finite.** Consider two regions A, B which do not touch each other. In the mutual information $I(A:B) = S(A) + S(B) - S(AB)$ the singular area-law terms in $S(A)$ and $S(B)$ cancel out. In particular, it has a chance to be finite (for finite-size subregions) in the continuum limit.

In the case where the whole state AB is pure, *i.e.* $B = \bar{A}$, we recover $I(A:B) = 2S(A)$ which has a UV-sensitive ('divergent') area law term. We can think of the mutual information as a regulator by considering a sequence of regions $B = \bar{A}^-$ which grow into $\bar{A}$ – the divergence occurs when their boundaries collide. For more on this point, see these papers of Casini and Huerta and Swingle.



• **Mutual information bounds correlations.** An important result is that mutual information gives a bound on correlation functions. Specifically, consider two regions of space A, B (perhaps separated by some distance), and any two operators $\mathcal{O}_A$ and $\mathcal{O}_B$ which act nontrivially only on A and B respectively – that is: $\mathcal{O}_A = M_A \otimes 1\!\!1_{\bar{A}}$ etc... (For example, $\mathcal{O}_A$ could be a local operator at some point in A.) Then

$$I(A:B) \geq \frac{1}{2} \frac{\langle \mathcal{O}_A \mathcal{O}_B \rangle_c^2}{\| \mathcal{O}_A \|^2 \| \mathcal{O}_B \|^2}. \tag{7.4}$$

Here the subscript on the correlator is for 'connected':

$$\langle \mathcal{O}_A \mathcal{O}_B \rangle_c \equiv \mathrm{tr} \boldsymbol{\rho} \mathcal{O}_A \mathcal{O}_B - \underbrace{\mathrm{tr} \boldsymbol{\rho} \mathcal{O}_A}_{= \mathrm{tr}_A \boldsymbol{\rho}_A \mathcal{O}_A} \cdot \underbrace{\mathrm{tr} \boldsymbol{\rho} \mathcal{O}_B}_{= \mathrm{tr}_B \boldsymbol{\rho}_B \mathcal{O}_B} = \mathrm{tr} \left(\boldsymbol{\rho} - \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B \right) \mathcal{O}_A \mathcal{O}_B .$$

The operator norms $\| X \| \equiv \sup \{ \sqrt{\langle \psi | X^\dagger X | \psi \rangle} \ s.t. \ \langle \psi | \psi \rangle = 1 \}$ in the denominator insure that the RHS doesn't change under rescaling the operators.

Here is a proof of (7.4) from [WVHC]: The mutual information is a relative entropy $I(A:B) = D(\boldsymbol{\rho}_{AB} || \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B)$. Relative entropy is bounded below in terms of the trace distance in the following way:

$$D(\boldsymbol{\rho} || \boldsymbol{\sigma}) \geq \frac{1}{2} \mathrm{tr} (\boldsymbol{\rho} - \boldsymbol{\sigma})^2 \tag{7.5}$$

[48] But then (twice) the RHS is of the form

$$\mathrm{tr} X^2 = \| X \|_2^2 \geq \| X \|_1^2 \quad ;$$

[48] (7.5) is Proposition 1.1 of the book by Ohya and Petz, *Entropy and its use.* It is also proved as eqn 11.22 of Petz' book *Quantum information theory and quantum statistics* which you can get electronically through the UCSD library here. The proof relies on convexity of $\eta(x) \equiv x \log x$ and the inequality $\eta(x) - \eta(y) + (x-y)\eta'(y) - \frac{1}{2}(x-y)^2 \geq 0$ for $x, y \in [0,1]$, with equality iff $x = y$. I mention this partly because this combination is just the combination that appears in the example I found (on the internet) of a function of two variables which is convex in both arguments but not jointly convex. There is some useful connection here that I am missing.

this last step is a special case of the Hölder inequality

$$\| XY \|_1 \le \| X \|_p \| Y \|_q, \qquad p^{-1}+q^{-1}=1 \tag{7.6}$$

for the Hilbert-Schmidt inner product, with $Y = \mathbb{1}$, X hermitian and $p = q = 2$. So, we have

$$D(\boldsymbol{\rho}_{AB}||\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B) \ge \frac{1}{2} \| \boldsymbol{\rho}_{AB} - \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B \|_1^2.$$

Now to get the operators in there on the RHS, we use the fact that

$$\| X \|_1 \ge \frac{\mathrm{tr} XY}{\| Y \|} \tag{7.7}$$

This is a Hölder inequality again with $p = 1, q = \infty$ – note that $\| X \| = \| X \|_\infty$. Taking $Y = \mathcal{O}_A \mathcal{O}_B$ our assumptions about their support means they commute and that $\| \mathcal{O}_A \mathcal{O}_B \| = \| \mathcal{O}_A \| \| \mathcal{O}_B \|$. $\boxed{7.4}$

So: here is a way in which this result can be used. At various points above I have used the thus-far-ill-defined term *correlation length*. If we are speaking about a collection of random variables Z_i distributed in space, this is usually defined as ξ in

$$\left\langle Z_x Z_y \right\rangle_c \overset{|x-y| \gg a}{\sim} e^{-|x-y|/\xi}$$

(up to power law prefactors). a is the lattice spacing. A power law means $\xi \to \infty$; if the correlations do not decay in this way, ξ isn't defined. In a general many-body system, there are many correlators to consider and ξ can depend on which we are talking about – some operators could be short-ranged, but others could be power laws. The mutual information bounds *all of them* and so provides an operator-independent definition of correlation length. If A and B are separated by distance r, $I(A : B) \overset{r \gg a}{\sim} e^{-r/\xi}$ says all other correlation lengths are bounded above by ξ.

Note that the bound (7.4) is not tight in general[49]. For example, the state could be a tensor product of N shared singlets between A and B, so the LHS is N. But the RHS is bounded above by 1/2.

Here are some area law-like statements about the mutual information in various many-body states.

• **Thermal classical states.** Consider a collection of $\mathfrak{D}$-states-per-site systems governed by $h = \sum_x h_x$ where each h_x is diagonal in some product basis (say the **Z**-basis). This means all the terms commute and further the groundstates are product states in the **Z** basis. The thermal state is $p(z) = e^{-\beta h(z)}/Z$, $(Z = \sum_z e^{-\beta h(z)})$ and in this case is best regarded as a probability distribution on the spins $\{z_i = 1...\mathfrak{D}\}$. This

[49] Thanks to T. Grover for discussions on this.

is a Markov chain in the sense that if two regions A and C are separated by a 'buffer region' B, so that no terms in h directly couple A and C, then

$$p(z_A|z_B z_C) = p(z_A|z_B). \tag{7.8}$$

The Markov property (7.8) implies $I(A:C|B) = 0$ when B separates A and C.

For two general regions then the mutual information is

$$I(A:B) = S(A) - S(A|B) = S(A) - S(A|B_0\partial B) = S(A) - S(A|\partial B) = I(A:\partial B) = I(\partial A:\partial B)$$

where ∂B is the set of sites in B directly connected to the exterior of B by terms in h, and B_0 is the rest of B. In the last step we used the same steps interchanging the roles of A and B to show that the answer is also independent of A_0. But then

$$I(A:B) = I(\partial A:\partial B) = H(\partial A) - \underbrace{H(\partial A|\partial B)}_{\geq 0 \quad \text{(classical!)}} \leq H(\partial A) \leq |\partial A| \log(\mathfrak{D})$$

where $\mathfrak{D}$ is the number of states per site, and $|\partial A|$ is the number of sites in the boundary region of A. So this is an area law. (The bound also holds with A replaced with B, so the one with the smaller boundary gives the stronger bound.) Notice that this result does not preclude power-law decay of correlations in thermal states, at second-order phase transitions, like the 2d Ising model at the critical temperature. The area law result doesn't tell us about the decay with distance between A and B.

- **Thermal quantum states.** Now consider thermal equilibrium $\boldsymbol{\rho} = \boldsymbol{\rho}_T = e^{-\beta\mathbf{H}}/Z$ for a general local Hamiltonian.

Lemma: For fixed $\mathbf{H}$ and fixed $T \equiv 1/\beta$, $\boldsymbol{\rho}_T = e^{-\beta\mathbf{H}}/Z$ minimizes the free-energy functional $F(\boldsymbol{\rho}) = \mathrm{tr}\boldsymbol{\rho}\mathbf{H} - TS(\boldsymbol{\rho})$ over all density matrices. Proof: for any state $\boldsymbol{\rho}$,

$$\begin{aligned} 0 \leq D(\boldsymbol{\rho}||\boldsymbol{\rho}_T) &= \mathrm{tr}\boldsymbol{\rho}\log\boldsymbol{\rho} - \mathrm{tr}\boldsymbol{\rho} \underbrace{\log\boldsymbol{\rho}_T}_{=-\beta\mathbf{H}-\log Z} \\ &= -S(\boldsymbol{\rho}) + \beta\mathrm{tr}\mathbf{H}\boldsymbol{\rho} + \log Z = \beta\left(\underbrace{\mathrm{tr}\mathbf{H}\boldsymbol{\rho} - S(\boldsymbol{\rho})}_{=F(\boldsymbol{\rho})} - \underbrace{T\log Z}_{-F(\boldsymbol{\rho}_T)}\right) \end{aligned} \tag{7.9}$$

So in particular for any subregion A, $F(\boldsymbol{\rho}_T) \leq F(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{\bar{A}})$ where $\boldsymbol{\rho}_A = \mathrm{tr}_{\bar{A}}\boldsymbol{\rho}_T$. This says

$$\mathrm{tr}\mathbf{H}\boldsymbol{\rho}_T - TS(\boldsymbol{\rho}_T) \leq \mathrm{tr}\mathbf{H}\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{\bar{A}} - T\left(S(A) + S(\bar{A})\right). \tag{7.10}$$

Now decompose the hamiltonian as $\mathbf{H}_A + \mathbf{H}_\delta + \mathbf{H}_{\bar{A}}$ where $\mathbf{H}_\delta$ contains all the terms which straddle the boundary between A and its complement. The terms in $\mathbf{H}_A$ are completely contained in A and $\mathrm{tr}\mathbf{H}_A\boldsymbol{\rho} = \mathrm{tr}_A\mathbf{H}_A\boldsymbol{\rho}_A$ and the same for $\bar{A}$. So, reorganizing (7.10) gives

$$\underbrace{S(A) + S(\bar{A}) - S(\boldsymbol{\rho}_T)}_{I(A:\bar{A})_{\boldsymbol{\rho}_T}} \leq \beta \mathrm{tr}\mathbf{H}_\partial \left(\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{\bar{A}} - \boldsymbol{\rho}_T\right) \leq 2\beta \| H_x \| |\partial A|$$

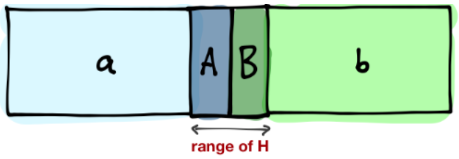
This is an area law for the mutual information in thermal states: the number of terms in the edge hamiltonian is $|\partial A| \sim R^{d-1}$. Notice that in the limit $T \to 0$, where the full system becomes pure, so that $I(A : \bar{A}) \to 2S(A)$, the RHS diverges and the bound goes away. So this does not prove a (false) area law for the EE without further assumptions.

- **Random singlets in 1d.** Specifically, consider a system of qbits on a line in a pure state of the following structure: For any given site i, the probability that i forms a singlet with another site j is $f(|i-j|)$ for some function f. This can be the groundstate of a Heisenberg antiferromagnet hamiltonian $H = \sum_{ij} J_{ij}\vec{S}_i \cdot \vec{S}_j$ with J_{ij} wildly varying in ij (but even with local J_{ij}, we can realize many examples of f). The entanglement entropy between a region and its complement is the number of singlets leaving the region. For a large region, this can be computed by averaging with the function $f(x)$, as can spin-spin correlation functions. This example is nice because the picture with the entanglement wormholes connecting the sites is actually literally correct.

7.3 Small incremental entangling by local Hamiltonians

Small Incremental Entangling Theorem (SIE) [Bravyi (conjectured by Kitaev, proved by van Acoleyen et al)]:

Consider a quantum system divided into parts $aABb$; the hamiltonian is local in the sense that only neighboring parts talk to each other directly.



Suppose the whole system is pure, and we will consider just the interactions between AB, so time evolution happens by

$$\mathbf{U} \equiv 1\!\!1_a \otimes e^{\mathbf{iH}_{AB}t} \otimes 1\!\!1_b$$

and a, b are regarded as ancillas.

Under the evolution by $\mathbf{U}(t)$, the EE of a pure state of $aABb$ satisfies

$$\partial_t S(Aa) \leq c\| \mathbf{H} \| \log \mathfrak{D}, \qquad \mathfrak{D} \equiv \min(|A|, |B|) \tag{7.11}$$

for some constant c independent of the sizes of the various Hilbert spaces. Notice that the coefficient on the RHS grows with the number of sites in the smaller of $|A|$ or $|B|$, not the dimension of the Hilbert space.

I will describe the argument for the special case where there are no ancillas a, b [from Bravyi]. Let the rate of change of entanglement be

$$\begin{aligned}\Gamma(\Psi, \mathbf{H}) \quad & \equiv \quad \partial_t S(A) \overset{\text{tr}\boldsymbol{\rho}_A=1}{=} -\text{tr}\dot{\boldsymbol{\rho}}_A \log \boldsymbol{\rho}_A \\ & \overset{\text{tr}[A,B]C \overset{\text{ibp}}{=} \text{tr}A[B,C]}{=} \mathbf{i}\text{tr}\mathbf{H} \underbrace{[\log \boldsymbol{\rho}_A \otimes \mathbb{1}_B, |\Psi\rangle\langle\Psi|]}_{\equiv X} \quad . \end{aligned} \tag{7.12}$$

This is linear in $\mathbf{H}$, so we can set $\|\mathbf{H}\| = 1$ and put it back at the end. For any Hermitian X, the Hölder inequality (7.7) says $\text{tr}(\mathbf{H}X) \leq \|\mathbf{H}\|\text{tr}|X|$ so that

$$\max_{\mathbf{H}|\|\mathbf{H}\|=1} \text{tr}\mathbf{H}X = \text{tr}|X| = \|X\|_1.$$

Therefore

$$\begin{aligned}\Gamma(\Psi) \equiv \max_{\mathbf{H}|\|\mathbf{H}\|=1} \Gamma(\Psi\mathbf{H}) &= \|[\log \boldsymbol{\rho}_A \otimes \mathbb{1}_B, |\Psi\rangle\langle\Psi|]\|_1, \qquad |\Phi\rangle \equiv \log \boldsymbol{\rho}_A \otimes \mathbb{1}|\Psi\rangle \\ &= \|\,|\Phi\rangle\langle\Psi| - |\Psi\rangle\langle\Phi|\,\|_1 \\ &= 2\sqrt{\langle\Psi|\Psi\rangle\langle\Phi|\Phi\rangle - |\langle\Psi|\Phi\rangle|^2} \equiv 2\sqrt{f(p)}. \end{aligned} \tag{7.13}$$

In the last step, we introduced the Schmidt decomposition of Ψ (the eigenvalues of $\boldsymbol{\rho}_A$):

$$|\Psi\rangle = \sum_{j=1}^{d} \sqrt{p_j}\,|j\rangle_A \otimes |j\rangle_B\,, \quad \boldsymbol{\rho}_A = \sum_j p_j\,|j\rangle\langle j|$$

and the function f is $f(p) \equiv \sum_{j=1}^{d} p_j \log^2 p_j - H(p)^2$ where $H(p)$ is the Shannon entropy. The evaluation of the trace norm can be done in the 2d (non-ON) basis spanned by $|\Phi\rangle, |\Psi\rangle$.

The dependence on Ψ of $\Gamma(\Psi)$ is thus all via the spectrum of $\boldsymbol{\rho}_A$, and finding the maximum is now a matter of calculus: $0 = \partial_{p_j} F(p) \implies -\log(2p_j) = H(p) \pm \sqrt{\ln^{-2}(2) + H(p)}$ which happens when $p = \left(\lambda, \underbrace{\frac{1-\lambda}{d-1}, \cdots, \frac{1-\lambda}{d-1}}_{d-1}\right)$ at which point

$$\frac{\Gamma(\Psi, \mathbf{H})}{\|\mathbf{H}\|} \leq \Gamma(\Psi) = 2\sqrt{f(p)} \leq 2\sqrt{\lambda(1-\lambda)} \log \frac{\lambda(d-1)}{1-\lambda} \leq c \log d.$$

I have not made it obvious that the ancillas a, b endanger the bound on the rate of entanglement, but indeed there are cases where they matter. Nevertheless, the van

Acoleyen paper proved (in a way that I haven't found it useful to try to reproduce here) that (7.11) continues to be true.

This result says that an area law is a property of a (gapped) phase. This is because within a gapped phase, by definition, the gap stays open. That means that there is an adiabatic path between any two representative Hamiltonians. Now apply the SIE theorem to the adiabatic time evolution. [50] [End of Lecture 19]

[50] More precisely, even with a uniform gap, the adiabatic evolution has some probability of producing an excited state that is nonzero per unit time and per unit volume. At the cost of slightly decreasing the locality of the time evolution operator, we can replace it by a 'quasilocal evolution' which is guaranteed to map groundstate to groundstate. This 'quasiadiabatic evolution' is a nice trick which Hastings explains here.

8 Quantum error correction and topological order

[Very readable is this review by Gottesman.] Earlier, I tried to convince you that quantum error correction would be difficult. Now I will convince you that it is possible.

Consider a noisy quantum channel which takes $|\psi\rangle \mapsto \mathsf{E}_i |\psi\rangle$ with probability p_i, with $\sum_i p_i \mathsf{E}_i^\dagger \mathsf{E}_i = \mathbb{1}$ (*i.e.* the Kraus operators are $\sqrt{p_i}\mathsf{E}_i$). This could be phase flip errors, *i.e.* decoherence, for example, if we take (on a single qbit)

$$\boldsymbol{\rho} \to (1-p)\boldsymbol{\rho} + p\mathbf{Z}\boldsymbol{\rho}\mathbf{Z}.$$

Recall that repeated action of this channel will erase the off-diagonal terms in $\boldsymbol{\rho}$. On the other hand, if we look at the same channel in the $\mathbf{X}$ basis, where $\mathbf{Z}|\pm\rangle = |\mp\rangle$, this acts as the classical binary symmetric channel. So bit flip and phase errors are canonically conjugate in this sense.

Suppose we can do the following encoding:

$$|0\rangle \mapsto |000\rangle \,, \quad |1\rangle \mapsto |111\rangle$$

to make a repetition code (this operation is linear and only acts as copy in the computational basis). We could then use majority rule to fix bit flip errors (in the $\mathbf{Z}$ basis). But phase flip errors would then be hopeless. Similarly, we could go to the $\mathbf{X}$ basis to do the repetition code ($|+\rangle \mapsto |+++\rangle \,, |-\rangle \mapsto |---\rangle$); then we could fix the phase flip errors (in the original basis), but then the bit flip errors would be hopeless.

It's possible to do both. Consider, for example, the following two 'code states' of 9 qbits:

$$|0\rangle \mapsto |0_L\rangle \equiv (|000\rangle + |111\rangle) \otimes (|000\rangle + |111\rangle) \otimes (|000\rangle + |111\rangle) \,.$$
$$|1\rangle \mapsto |1_L\rangle \equiv (|000\rangle - |111\rangle) \otimes (|000\rangle - |111\rangle) \otimes (|000\rangle - |111\rangle) \,.$$

This is Shor's 9-qbit code. When I have to, I will label the qbits Z_{xy} where $x = 1,2,3$ indicates which group of three it lives in and $y = 1,2,3$ is which member of the group, but for now let's distinguish them by their lexicographic position. Consider the following Hamiltonian :

$$\begin{aligned}-\mathbf{H} = &(ZZ1)(111)(111) + (111)(ZZ1)(111) + (111)(111)(ZZ1) + (XXX)(XXX)(111) \\ &+ (1ZZ)(111)(111) + (111)(1ZZ)(111) + (111)(111)(1ZZ) + (111)(XXX)(XXX)\end{aligned}$$

The terms in $\mathbf{H}$ (called *stabilizers*) all commute with each other, and further, both code states $|a_L\rangle \,, (a = 0,1)$ are eigenstates with smallest possible eigenvalue (-1 for every term in $\mathbf{H}$).

It is useful to denote this in the same way as we did for Hamming codes: each row is associated with a stabilizer – the 1s above the line indicate where the Zs go, and the ones below indicate where the Xs go. Since they all commute, the coefficients don't matter, as long as they are positive. The only thing that matters (for the groundstates) is the *algebra* generated by multiplying (and adding with positive coefficients) the stabilizers. In particular, we could include *e.g.* $(1ZZ)(111)(111)$ and $(111)(XXX)(XXX)$ without changing anything.

$$G \equiv \left(\begin{array}{ccccccccc} 1&1&0&0&0&0&0&0&0\\ 0&1&1&0&0&0&0&0&0\\ 0&0&0&1&1&0&0&0&0\\ 0&0&0&0&1&1&0&0&0\\ 0&0&0&0&0&0&1&1&0\\ 0&0&0&0&0&0&0&1&1\\ \hline 1&1&1&1&1&1&0&0&0\\ 0&0&0&1&1&1&1&1&1 \end{array}\right)$$

Notice that $\mathbf{X}_L \equiv \mathbf{Z}_{1y}\mathbf{Z}_{2y}\mathbf{Z}_{3y}$ (for any y) flips between $|0_L\rangle$ and $|1_L\rangle$. It acts as a 'logical X' operator. Similarly, the two states $|0_L\rangle$ and $|1_L\rangle$ are eigenstates of $\mathbf{Z}_L \equiv \mathbf{X}_{x1}\mathbf{X}_{x2}\mathbf{X}_{x3}$ with eigenvalues ± 1 respectively. These operators anticommute with each other (they share the one $\mathbf{X}$ and $\mathbf{Z}$ at position xy) but commute with the whole stabilizer algebra.

Errors. We can check for bit flip errors by measuring the stabilizers with Zs, just like majority rule. And this can be done without messing with the state: introduce an ancilla qbit, initially in a product state, and then act with a unitary which is $ZZ1$ controlled by 1 (*i.e.* $C_{ZZ1} \equiv (ZZ1)^{\frac{1}{2}(1-\mathbf{Z}_{\text{ancilla}})}$):

$$(|0\rangle + |1\rangle) \otimes \sum_{abc=0,1} \psi_{abc} |abc\rangle \overset{C_{ZZ1}}{\to} \sum_{abc} \psi_{abc} \left(|0\rangle |abc\rangle + |1\rangle ZZ1 |abc\rangle\right) = \sum_{abc} \psi_{abc} \left(|0\rangle + |1\rangle (-1)^{(a+b)_2}\right) |abc\rangle$$

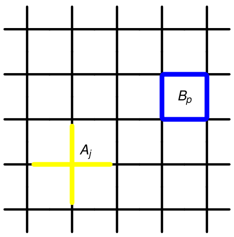
and then measure the $X_{\text{ancilla}} = \pm 1$. If you get (-1) it means there was an odd number of flips of a, b (which means one flip since they are $\mathbb{Z}_2$-valued).

But now (unlike the repetition code), we can also check for sign flips by measuring the stabilizers which involve X, too (since they commute). So Shor's code is a 9 qubit code which encodes 1 logical qubit and is robust to any single qubit error. I found it very hard to keep in my head until I learned the following amazing generalization.

8.1 Toric code

First, here's the toric code. It's a paradigmatic example of a system with topological order. It emerges $\mathbb{Z}_2$ gauge theory from a local Hilbert space.

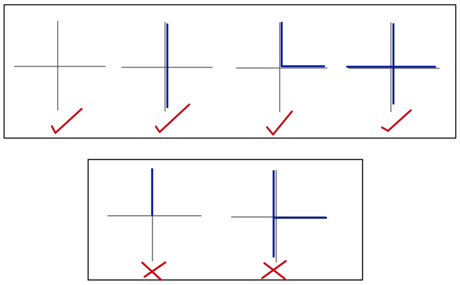
Consider a 2d simplicial complex. This means a graph (a set of vertices who know with whom they share an edge) with further information about plaquettes, who know which edges bound them). For example, consider the square lattice at right. Now place a qubit on each *edge*. Now let's make some stabilizers. Associate to each plaquette a 'flux operator', $B_p = \prod_{\ell \in p} Z_\ell$, and to each vertex a 'gauss law operator', $A_v = \prod_{\ell \in v} X_\ell$. (These names are natural if we consider Z to be related to a gauge field by $Z \sim e^{\mathbf{i}A}$, and X is its electric flux. For more on the translation to gauge theory see §5.2 here.)



[Fig by D.Ben-Zion, after Kitaev]

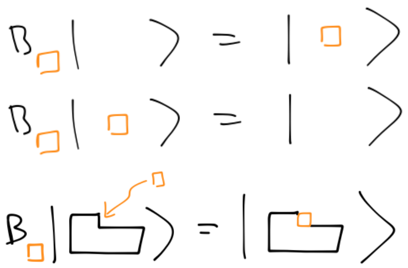
The hamiltonian is $\mathbf{H}_{\text{TC}} = -\Gamma_m \sum_p B_p - \Gamma_e \sum_v A_v$. These terms all commute with each other (since each vertex and plaquette share zero or two links), and they each square to one, so the Hamiltonian is easy to diagonalize.

Which states satisfy the 'gauss law condition' $A_v = 1$? In the X basis there is an extremely useful visualization: we say a link l of $\hat{\Gamma}$ is covered with a segment of string (an electric flux line) if $\mathbf{e}_l = 1$ (so $X_l = -1$) and is not covered if $\mathbf{e}_l = 0$ (so $X_l = +1$): $\equiv X = -1$. In the figure at right, we enumerate the possibilities for a 4-valent vertex. $A_v = -1$ if a flux line ends at v.



So the subspace of $\mathcal{H}$ satisfying the gauss law condition is spanned by closed-string states (lines of electric flux which have no charge to end on), of the form $\sum_{\{C\}} \Psi(C) |C\rangle$.

Now we look at the action of B_p on this subspace of states:



$B_p = \prod_{\ell \in \partial p} Z_\ell$ creates and destroys strings around the boundary of the plaquette:

$$B_p |C\rangle = |C + \partial p\rangle \ .$$

The condition that $B_p |\text{gs}\rangle = |\text{gs}\rangle$ is a homological equivalence. In words, the eigenvalue equation $\mathbf{B}_\square = 1$ says $\Psi(C) = \Psi(C')$ if C' and C can be continuously deformed into each other by attaching or removing plaquettes.

If the lattice is simply connected – if all curves are the boundary of some region contained in the lattice – then this means the groundstate

$$|\text{gs}\rangle = \sum_C |C\rangle$$

is a uniform superposition of all loops.

Topological order. If the space has non-contractible loops, however, then the eigenvalue equation does not determine the relative coefficients of loops of different topology! On a space with $2g$ independent non-contractible loops, there are 2^{2g} independent groundstates.

No local operator mixes these groundstates. This makes the topological degeneracy stable to local perturbations of the Hamiltonian. The degenerate groundstates are instead connected by the action of (Wilson) loop operators:

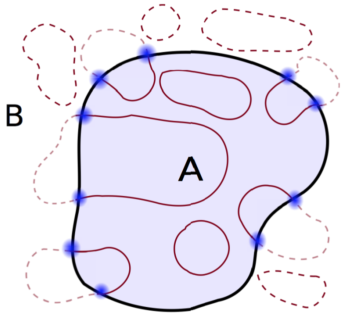
$$W_C = \prod_{\ell \in C} Z_\ell \quad V_{\check{C}} = \prod_{\ell \perp \check{C}} X_\ell \ .$$

(Notice that the loop operator for a single plaquette $W_{\partial\Box} = B_p$ is the plaquette operator.) V, W commute with $\mathbf{H}_{\text{TC}}$ and don't commute with each other (specifically W_C *anti*commutes with $V_{\check{C}}$ if C and $\check{C}$ intersect an odd number of times). This algebra must be represented on the groundstates, and it doesn't have any one-dimensional representations.

Here is why **deconfined flux lines** (the fact that the groundstate is a superposition of arbitrarily large closed loops) mean **long-range entanglement**: The picture at right shows why – in a state described by fluctuating *closed* strings – there is a contribution to the EE of region A which is independent of the size of A: if a string enters A and it is known to be closed, then it must leave again somewhere else; this is one missing bit of freedom, so

$$S = \frac{R}{\epsilon} - \log 2 \equiv \frac{R}{\epsilon} - \gamma \tag{8.1}$$

where the area-law coefficient ϵ is some short-distance-dependent junk and γ is a universal characterization of the nature of the topological order. More on this in §8.2.

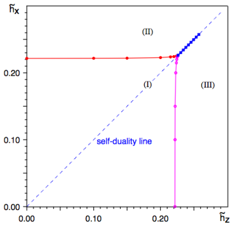


[fig: Tarun Grover]

8.1.1 The deconfined phase

So far everything I've said works on any graph (actually: cell complex). And so far I've described the solvable limit. But this physics actually pertains to an open subset of the phase diagram – the toric code is a representative of a phase of matter.

Perturbations $\Delta H = \sum_l (hX_l + gZ_l)$ produce a nonzero correlation length. Let's focus on $D = 2+1$ for what follows. These couplings h and g are respectively a string tension and a fugacity for the electric flux string endpoints: charges. Make these too big and the model is confined or higgsed, respectively. These are actually adiabatically connected [Fradkin-Shenker]: Both are connected to the trivial state where e.g. $H = \sum_l X_l$ whose groundstate is a product $\otimes_l |{\to}_l\rangle$.



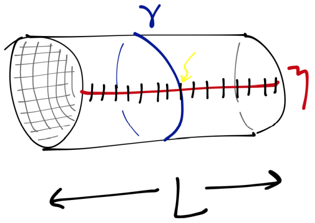
[from Tupitsyn-Kitaev-Prokof'ev-Stamp]

Defects. There are two kinds of elementary excited states of the toric code: violations of $A_s = 1$ and violations of $B_p = 1$. [51] The defects are created by the endpoints of open Wilson lines. Again there are two kinds:

$$W(C) = \prod_{\ell \in C} Z_\ell, \quad V(\check{C}) = \prod_{\ell \perp \check{C}} X_\ell.$$

Here C is an open curve in the lattice, and $\check{C}$ is an open curve in the dual lattice. Endpoints of $W(C)$ violate A_s and endpoints of $V(\check{C})$ violate B_p.

Consider the cylinder. There is one nontrivial class of loops; call a representative γ. Let η be a line running along the cylinder. The two groundstates are generated by the action of the Wilson loop operator



$$V(\eta) \equiv \prod_{\ell \text{ crossed by } \eta} X_\ell$$

in the sense that

$$|\text{gs}_2\rangle = V(\eta)\,|\text{gs}_1\rangle \ .$$

This is also a groundstate (at $g, h = 0$) since there is no plaquette which violates $\mathbf{B}_p = 1$ (more simply: $[\mathbf{H}_{g=h=0}, V_x(\eta)] = 0$). They are *distinguished* by $W(\gamma) \equiv \prod_{l\in\gamma} X_l$ in the sense that the two groundstates are eigenstates of this operator with distinct eigenvalues:

$$W(\gamma)\,|\text{gs}_\alpha\rangle = (-1)^\alpha\,|\text{gs}_\alpha\rangle\,, \quad \alpha = 1, 2.$$

This follows since $W(\eta)V(\gamma) = -V(\gamma)W(\eta)$ – the two curves share a single link (the one pointed to by the yellow arrow in the figure).

[51] Cultural note: The limit where the coefficient of the star term A_s goes to infinity is called 'pure $\mathbb{Z}_2$ gauge theory'. The condition $A_s = 1$ is the called Gauss' law constraint. The former kind of defect would cost infinite energy and hence are strictly forbidden in this theory.

At finite g, h (and in finite volume), there is tunneling between the topologically degenerate groundstates, since in that case

$$[\mathbf{H}, \prod_{l \in \gamma} X_l] \neq 0.$$

This means that for some n

$$\langle \mathrm{gs}_2 | \mathbf{H}^n | \mathrm{gs}_1 \rangle \neq 0.$$

The process which mixes the groundstates requires the creation of magnetic flux on some plaquette (*i.e.* a plaquette P with $B_P = -1$, which costs energy $2\Gamma_m$), which then must hop (using the hX term in $\mathbf{H}$) all the way along the path η, of length L, to cancel the action of $V(\eta)$. The amplitude for this process goes like

$$\Gamma \sim \frac{\langle \mathrm{gs}_2 | (hX_1)(hX_2)\cdots(hX_L) | \mathrm{gs}_1 \rangle}{2\Gamma_m \cdot 2\Gamma_m \cdot \ldots 2\Gamma_m} \sim \left(\frac{h}{2\Gamma_m}\right)^L = e^{-L|\ln 2\Gamma_m/h|}$$

which is *extremely tiny* in the thermodynamic limit. The way to think about this is that the Hamiltonian is itself a local operator, and cannot distinguish the groundstates from each other. It takes a non-perturbative process, exponentially suppressed in system size, to create the splitting.

8.2 Entanglement, short and long

Mean field theory is product states, which means there is no entanglement between regions of space at all. The next level of complication and interest to consider for possible groundstates of quantum many body systems is the case of states obtained by acting with a short-ranged quantum circuit of small depth on a product state. Let us consider such states, which are called short-range-entangled. What does their entanglement entropy of subregions look like and how do we distinguish which bits might be properties of a phase?

Consider $d = 2$. If the entanglement is short-ranged, we can construct a local 'entanglement entropy density' which is supported along the boundary of the region A [Grover-Turner-Vishwanath]:

$$S(A) = \oint_{\partial A} s d\ell = \oint \left(\Lambda + bK + cK^2 + ...\right) = \Lambda \ell(\partial A) + \tilde{b} + \frac{\tilde{c}}{\ell(\partial A)} + ...$$

In the first step, we use the fact that the entanglement is localized at the boundary between the region and its complement. In the second step we parametrize the local entropy density functional in a derivative expansion; K is the extrinsic curvature of the boundary. Since the total system is in a pure state, $S(A) = S(\bar{A})$, which implies $b = 0$:

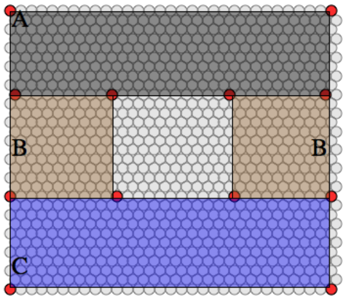
since interchanging A and $\bar{A}$ reverses the orientation of the boundary, the extrinsic curvature cannot contribute. This means that the subsystem-size-independent term cannot come from terms local on the boundary; it is universal, and cannot be changed by changing the UV regulator.

If the entanglement is indeed all short-ranged, then SSA is saturated for collections of regions where the boundaries cancel out, $\partial(AB) + \partial(BC) = \partial(B) + \partial(ABC)$, such as in the figure below.

Let $S(x)$ be the EE of the subregion x in the state in question.

$$I(A : C|B) := S(AB) + S(BC) - S(B) - S(ABC)$$

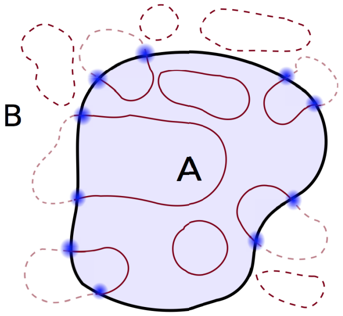
is the conditional mutual information – correlations between variables A and C if we knew B. In general gapped phases in 2d, for the arrangement of regions at right, $I(A : C|B) = 2\gamma$, where the generalization of (8.1) to regions with $b_0(\partial A)$ boundary components is $S(A) = |\partial A|\Lambda - \gamma b_0(\partial A)$. The area-law contributions cancel out pairwise (notice that the corners cancel too). $\gamma \geq 0$ by SSA.



So the deviation from saturation of SSA here (which is called, naturally, the topological entanglement entropy (TEE)) is a diagnostic for long-ranged entanglement. A term in the EE which would produce a nonzero TEE is a constant, independent of the size of the region. In such a state $S(A) = \Lambda\ell(\partial A) - \gamma$ (Λ is the UV cutoff on wavenumber). The *deficit* relative to area law, γ, is called the "topological entanglement entropy" [52].

Why a deficit relative to the area law? Again, for the example of the groundstate of $\mathbb{Z}_2$ gauge theory (the toric code), a closed string that enters a region must leave again. This is true for each component of the boundary individually.

More generally, (for Abelian states) it is proportional to the log (#torus groundstates) ≥ 0. A beautiful argument for this is the Kitaev-Preskill wormhole construction (see their Fig. 2).



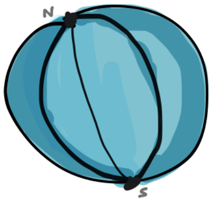
[fig: Tarun Grover]

When $I(A : C|B) = 0$ it means $\boldsymbol{\rho}_{ABC}$ is a quantum Markov chain and can be reconstructed from the marginals (by the Petz formula, described very briefly below). The nonvanishing quantity γ is an obstruction to this automatic reconstruction of the state from local data.

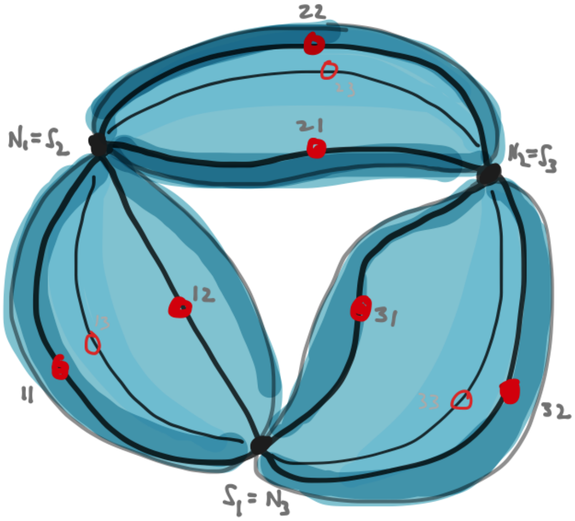
[52] introduced for $d = 2$ by Hamma-Ionicioiu-Zanardi, Kitaev-Preskill, Levin-Wen; the higher-dimensional generalizations are clarified in the Grover et al paper linked above.

8.3 Shor's code is a toric code

The following beautiful thing was explained to me by Brian Swingle: Shor's code is $\mathbb{Z}_2$ gauge theory on a certain complex.



The complex is constructed by taking a two-sphere, marking the north and south poles (N and S), and connecting the north to the south pole by three edges. These three edges break the sphere into three orange slices.

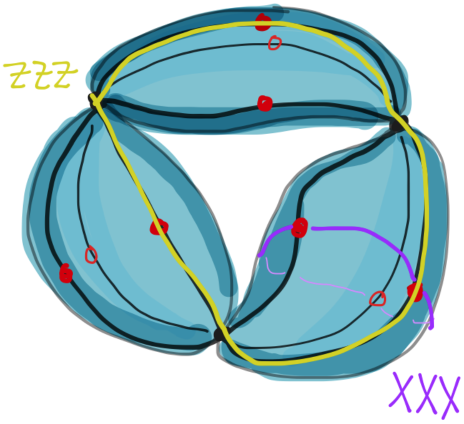


Now take three such spheres and glue N_1 to S_2, N_2 to S_3, and N_3 to S_1, therby making a closed chain. The resulting object has 9 edges (3 edges per sphere), 3 vertices, and 9 faces (3 faces per sphere). The resulting space has one non-contractible loop, going around the chain of spheres.

Now (surely you saw this coming): put the toric code on this complex. There are three vertices. The star terms (of which there are three) each involve six Xs, three from each of two neighboring spheres. The algebra is generated by just two of them.

The plaquette terms (of which there are 9) each involve two Zs from links bounding the same segment of orange peel. Two of the three pairs from a given orange multiply to give the third.

Its ground state is two-fold degenerate and is the code subspace of Shor's code!



The logical operators are the Wilson line which wraps the chain of three spheres (the yellow ZZZ in the figure at right), and the conjugate string operator made of (any) three Xs from a single sphere (purple XXX). Different choices of path differ by terms in the hamiltonian, which act as the identity on the code subspace. [End of Lecture 20]

9 Tangent vectors to an imagined future

Here I will briefly summarize some natural next steps which we will not have time to take together, *i.e.*, some of the many other ways in which ideas from quantum information theory can be useful in thinking about quantum many body systems.

When is there an area law? There are some cases where the area law is a rigorous statement. Hastings' 1d area law theorem proves that the area law is true for groundstates of one-dimensional local Hamiltonians with an energy gap, and hence that there is a good MPS representation for such states.

The theorem was proved using the...

Lieb-Robinson bound. Even non-relativistic theories have lightcones. Given a local Hamiltonian $\mathbf{H} = \sum_Z H_Z$ where the terms H_Z are supported on a subset Z and $\| H_Z \|$ shrinks rapidly with the diameter of Z (exponentially is good), then we can bound the correlations of local operators (A_X is supported on a set X and $A_X(t)$ is its time evolution by $\mathbf{H}$):

$$\| [A_X(t), B_Y] \| \leq c e^{-a d_{XY}} \left(e^{2st} - 1 \right)$$

where $d_{XY} = \min_{i \in X, y \in Y} |i - j|$ is the distance between the sets X, Y and $c = 2\| A_X \| \| B_Y \| |X|$ is a constant. The quantity $2s/a$ is the *Lieb-Robinson velocity.*

The ocean of volume law states. Consider $\mathcal{H} = \mathcal{H}_m \otimes \mathcal{H}_n, m \leq n$. Let us associate these factors with regions of space A and $\bar{A}$, so that

$$\log(m) = \log d_{\text{local}} \times (\#\text{of sites in region A}) \propto \text{Volume}(A).$$

Let us consider a *random* state $|w\rangle \in \mathcal{H}$: $|w\rangle = \mathbf{U} |w_0\rangle$ for some reference state $|w_0\rangle$ and $\mathbf{U}$ is chosen from the Haar measure on $\mathsf{U}(mn)$. How entangled is such a state, on average? The answer is: almost as entangled as possible, *i.e.* volume law: $S \propto \text{Volume}(A)$.

Here's a sketch of the calculation: The von Neumann entropy of the subsystem A depends only on the eigenvalues of the reduced density matrix. So we can do most of the integrals $\int d^{(nm)^2}\mathbf{U}$ in the Haar measure, and their only effect is to change the measure for the eigenvalues λ of $\boldsymbol{\rho}_A$.

$$\begin{aligned} \langle S(\boldsymbol{\rho}_A) \rangle &= \int \prod_{i=1}^m d\lambda_i P_{m,n}(\lambda) S(\lambda) \\ &= \prod_{i=1}^m d\lambda_i C_{mn} \delta \left(\sum_i \lambda_i - 1 \right) \prod_i \lambda_i^{n-m} \prod_{i<j} \left(\lambda_i - \lambda_j \right)^2 S(\lambda) \end{aligned} \quad (9.1)$$

where the normalization factor is basically a multinomial $C_{mn} = \frac{\Gamma(mn)}{\prod_{i=0}^{m-1}\Gamma(n-i)\Gamma(m-i+1)}$. This integral can be done exactly, but the limit of $m \gg n$ gives

$$\langle S(\boldsymbol{\rho}_m)\rangle = \log m - \frac{m}{2n} + \ldots$$

(This limit is relevant when the subsystem is a small fraction of the whole system.) This is sometimes called *Page's theorem*, although Page wasn't quite the last to prove it. So a thermal state is just as entangled as a completely random state. Didn't we prove that most of these states are unreachable by physical systems?

Eigenstate thermalization. I didn't say enough about eigenstate thermalization. In case you missed it, look at footnote 36.

Bekenstein bound. The positivity of the relative entropy implies of version of the fabled Bekenstein bound $S \leq \frac{2\pi}{\hbar c} RE$ where, roughly, S is the entropy of a system, R is its linear size and E is its energy. This relation was argued by Bekenstein by demanding a consistent extension of thermodynamics in the presence of black holes, but the relation itself does not involve gravity. A precise version was shown in this paper by Casini. I mentioned above (in our discussion of 'entanglement thermodynamics' in §5) that the entanglement hamiltonian for a half-line in a relativistic QFT is the boost generator, $\int dx\, xH_x$; this is how the RHS arises. The danger of adding many species of particles (which seems to grow the LHS but not the RHS of the Bekenstein inequality) is resolved by the joint convexity of the relative entropy!

Renormalization group monotones from entanglement measures. The renormalization group (RG, the subject of Physics 217 in the Fall) is a procedure by which degrees of freedom are thinned; starting from a microscopic theory of all degrees of freedom, it is possible to coarse-grain our description in order to obtain a (different) theory of just the long-wavelength degrees of freedom. This procedure is hard to do in practice, and it is useful to know about quantities which behave monotonically under this process; just a quantity is then naturally regarded as a measure of the number of degrees of freedom.

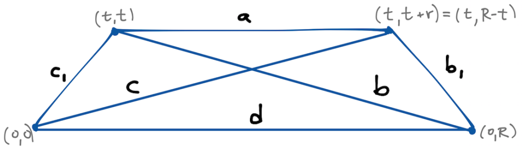
The quantity c appearing in (7.2), the entanglement entropy of a single interval in the groundstate of a 1+1d CFT is such a quantity. It was proved to be an RG monotone by Zamolodchikov long ago. A proof of its monotonicity using Lorentz invariance and SSA was found by Casini and Huerta. (This really belongs in §5 on many-body applications of SSA.)

Proof of entropic c-theorem in $d = 1 + 1$. We will show that the entanglement entropy of an interval of length r satisfies

$$0 = \left(\frac{d}{d\log r}\right)^2 S(r). \tag{9.2}$$

This says that $c(r) \equiv r\partial_r S(r)$ satisfies $c'(r) = rS'' - S' \leq 0$. In particular, if $S(r) = \frac{c}{3}\log r/\epsilon$, then $c(r) = c$ is the central charge.

First a Minkowski space geometry exercise. Consider an interval d, and the inward light rays from its endpoints b_1, c_1. Extend these for a time t. The interval connecting the endpoints of the lightrays is a. Let b and c be the spacelike surfaces in the figure. Lorentzian geometry then implies $bc = ad$, where the letters indicate the proper lengths of the associated segments. This says $c = \lambda a, d = \lambda b, \lambda = \frac{c}{a} = \frac{d}{b} \geq 1$.



More explicitly, if $d \equiv |d^\mu d_\mu| = R, a = r$, then $c^\mu = (t, t+r)^\mu = (t, R-t)^\mu$ has $c^2 = |c^\mu c_\mu| = r(r-2t) = rR = |b^2| = bc$.

SSA says

$$\underbrace{S(c_1 a)}_{=S(c)} + \underbrace{S(ab_1)}_{=S(b)} \geq S(a) + S(d).$$

The underbraced equations are consequences of Lorentz invariance. Then

$$S(b) - S(a) \geq S(\lambda b) - S(\lambda a).$$

Notice that the area law term cancels in the differences. If we set $\lambda = 1 + \epsilon$, (9.2) follows.

In $D = 3, 4$, universal terms in the entanglement entropy of subregions have also been shown to be RG monotones. This paper has references which explain this long story.

Recovery and reconstruction. Cover space with overlapping patches A_i. Take a state $\boldsymbol{\rho}$ on the whole space and let $\boldsymbol{\rho}_i \equiv \mathrm{tr}_{\overline{A_i}}\boldsymbol{\rho}$ be the reduced states. The existence of a global state implies consistency conditions between the reduced states on the patches when they intersect

$$\boldsymbol{\rho}_{ij} \equiv \mathrm{tr}_{\overline{A_i \cap A_j}}\boldsymbol{\rho} = \mathrm{tr}_{\overline{A_i \cap A_j} \subset A_i}\boldsymbol{\rho}_i \overset{!}{=} \mathrm{tr}_{\overline{A_i \cap A_j} \subset A_j}\boldsymbol{\rho}_j.$$

The other direction is much harder: Determining a density matrix from its marginals is not simple[53] for just the reasons that SSA of quantum entropy is hard. In fact, there are some consistent density matrices which do not permit any global state: for example, if $\boldsymbol{\rho}_{12}$ is pure, then $\boldsymbol{\rho}_{23} = \mathrm{tr}_1\boldsymbol{\rho}_{123} = \mathrm{tr}_1(\boldsymbol{\rho}_{12} \otimes \boldsymbol{\rho}_3) = \boldsymbol{\rho}_2 \otimes \boldsymbol{\rho}_3$ must factorize. Here are some references: Carlen-Lebowitz-Lieb, Swingle-Kim. The latter can be regarded as a generalization of density functional theory.

[53] unlike the classical case, where Bayes' formula gives a preferred answer $p(123) = \frac{p(12)p(23)}{p(2)}$ which by SSA maximizes the entropy $S(12) + S(23) - S(2)$ over all possible reconstructions

The above 'quantum marginals' problem is a special case of the problem of reversing a quantum channel (for the special case of partial trace). There is a general solution of this problem, adapted to a particular input, called the *Petz recovery channel*: given a channel $\mathcal{E}$ from A to B and a particular state $\boldsymbol{\sigma}$ on A, there exists a channel $\mathcal{R}_{\mathcal{E},\boldsymbol{\sigma}}$ from B to A such that

$$\mathcal{R}_{\mathcal{E},\boldsymbol{\sigma}}(\mathcal{E}(\boldsymbol{\sigma})) = \boldsymbol{\sigma} \tag{9.3}$$

It's simple: If $\{\mathcal{M}_k\}$ are Kraus operators for $\mathcal{E}$, then the Kraus operators for $\mathcal{R}_{\mathcal{E},\boldsymbol{\sigma}}$ are $\{\sqrt{\boldsymbol{\sigma}}\mathcal{M}_k^\dagger \mathcal{E}(\boldsymbol{\sigma})^{-1/2}\}$ (where as usual the inverse is defined on the image of the map). Check that the resulting channel is trace-preserving and positive and achieves (9.3).

What it does to other states we don't answer. But under some circumstances, one can appeal to a version of typicality to use this to approximately invert other states.

This map gives a useful statement (due to Petz) of when monotonicity of the relative entropy is saturated: $D(\boldsymbol{\rho}||\boldsymbol{\sigma}) \geq D(\mathcal{E}(\boldsymbol{\rho})||\mathcal{E}(\boldsymbol{\sigma})$ with equality IFF $\exists$ a channel $\mathcal{R}$ from B to A such that $\mathcal{R} \circ \mathcal{E}(\boldsymbol{\rho}) = \boldsymbol{\rho}$ and $\mathcal{R} \circ \mathcal{E}(\boldsymbol{\sigma}) = \boldsymbol{\sigma}$ (with the same map). When it exists, it is the Petz map.

A strengthening of SSA [due to Fawzi and Renner] and of the monotonicity of the relative entropy constitute a frontier of recent progress. In particular, they can put a positive something on the RHS where SSA has a zero:

$$I(A:C|B)_{\boldsymbol{\rho}} \geq D_{\mathcal{M}}(\boldsymbol{\rho}_{ABC}||\mathcal{R}_{B\to BC}(\boldsymbol{\rho}_{AB})) \ .$$

For future reference, $\mathcal{R}$ is the Petz recovery channel for the partial trace:

$$\mathcal{R}_{B\to BC} : X_B \to \mathbf{V}_{BC}\sqrt{\boldsymbol{\rho}_{BC}}\left(\boldsymbol{\rho}_B^{-1/2}\mathbf{U}_B X_B \mathbf{U}_B^\dagger \otimes \mathbb{1}_C\right)\sqrt{\boldsymbol{\rho}_B}\mathbf{V}_{BC}^\dagger$$

and $D_{\mathcal{M}}$ is made from the relative entropy by

$$D_{\mathcal{M}}\left(\boldsymbol{\rho}||\boldsymbol{\sigma}\right) \equiv \sup_{\text{POVMs},\mathcal{M}} \{D\big(\mathcal{M}(\boldsymbol{\rho})||\mathcal{M}(\boldsymbol{\sigma})\big) \ \big| \ \mathcal{M}(\boldsymbol{\rho}) = \sum_x \left(\text{tr}\boldsymbol{\rho}\mathcal{M}_x\right)|x\rangle\langle x| \, , \sum_x \mathcal{M}_x = \mathbb{1}\}.$$

In particular Brian Swingle and I are using these results as part of a program which we call...

***s*-sourcery.** This is a scheme for hierarchical growth of entangled states. It gives a demonstration of the area law for a large class of states, and suggests some new routes to efficient constructions of their wavefunctions.

Finite-temperature quantum memory. The toric code is great, but at any nonzero temperature there is a finite density of violated stabilizers, and under generic perturbations of $\mathbf{H}_{\text{TC}}$, these become mobile and mix the code states. The version with the qbits living on the plaquettes of a 4d lattice does better, but the very interesting state of the art in 3 or fewer dimensions falls short so far.

Solving local Hamiltonians is a Hard® problem. If you can efficiently find the groundstate of any 1d Hamiltonian with $d_L = 12$ you can solve any problem in QMA (roughly: any problem whose answer you can check and which is difficult for a quantum computer). See this paper. These models are somewhat artificial, but more generally, the connection between Hard Problems and quantum many-body groundstates continues to be interesting; see, for example, this and this.

Apology about the thermodynamic limit. Our stated motivation in this course has been the application of ideas from information theory and quantum information theory to many-body physics. This differs from the general problem in two ways: First, it implies that we have a notion of locality, which is a simplification we've incorporated at every opportunity. In fact, you could say that the job of understanding the implications of locality is the main subject here.

Second, as we argued at the beginning, the most interesting physics happens in the thermodynamic limit when the number of degrees of freedom goes to infinity. Sometimes we've also been interested in a continuum limit, where the number of sites per unit volume also diverges. In both cases, the dimension of the Hilbert space is infinite. Given this, I have to admit that I have been somewhat remiss in not being more careful about which results, and perhaps more importantly which techniques, can break down for infinite dimensional Hilbert spaces.

Exercise: go back over everything we've learned and see which statements actually depend on finite dimension of $\mathcal{H}$.

References for the table of entanglement measures

[1] M. Christandl and A. Winter. “Squashed entanglement” - an additive entanglement measure. *J. Math. Phys.*, 45:829, 2004. 126

[2] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters. Mixed state entanglement and quantum error correction. *Phys. Rev. A,*, 54:3824, 1996. 126

[3] E. M. Rains. Rigorous treatment of distillable entanglement. *Phys. Rev. A*, 60:173, 1999. 126

[4] I. Devetak and A. Winter. Distillation of secret key and entanglement from quantum states. *Proc. Roy. Soc. Lond. Ser. A*, 461:207, 2004. 126

[5] K. Horodecki, M. Horodecki, P. Horodecki, and J. Oppenheim. Secure key from bound entanglement. *Phys. Rev. Lett.*, 94:160502, 2005. 126

[6] P. Hayden, M. Horodecki, and B. M. Terhal. The asymptotic entanglement cost of preparing a quantum state. *J. Phys. A,*, 34:6891, 2001. 126

[7] V. Vedral, M. B. Plenio, M. A. Rippin, and P .L. Knight. Quantifying entanglement,. *Phys. Rev. Lett.*, 78:2275, 1997. 126

[8] V. Vedral and M. B. Plenio. Entanglement measures and purification procedures. *Phys. Rev. A*, 57:1619, 1998. 126

[9] G. Vidal and R. F. Werner. A computable measure of entanglement. *Phys. Rev. A*, 65:032314, 2001. 126

[10] M. Horodecki, P. Horodecki, and R. Horodecki. Mixed-state entanglement and distillation: Is there a ”bound” entanglement in nature? *Phys. Rev. Lett.*, 80:5239, 1998. 126

[11] D. Yang, M. Horodecki, R. Horodecki, and B. Synak-Radtke. Irreversibility for all bound entangled states. *Phys. Rev. Lett.*, 95:190501, 2005. 126

[12] F. G. S. L. Brandão and M. B. Plenio. A generalization of quantum Stein’s lemma. *Comm. Math. Phys.*, 295:791, 2010.

[13] M. Christandl. *The Structure of Bipartite Quantum States - Insights from Group Theory and Cryptography*. PhD thesis, Cambridge University, 2006. 126

[14] M. Christandl, N. Schuch, and A. Winter. Highly entangled states with almost no secrecy. *Phys. Rev. Lett.*, 104:240405, 2010. 126

[15] M. Koashi and A. Winter. Monogamy of entanglement and other correlations. *Phys. Rev. A*, 69:022309, 2004. 126

[16] M. B. Plenio. Logarithmic negativity: A full entanglement monotone that is not convex. *Phys. Rev. Lett.*, 95:090503, 2005. 126

[17] R. Alicki and M. Fannes. Continuity of quantum conditional information. *J. Phys. A: Math. Gen.*, 37:L55, 2004. 126

[18] M. J. Donald and M. Horodecki. Continuity of relative entropy of entanglement. *Physics Letters A*, 264:257, 1999. 126

[19] M. J. Donald, M. Horodecki, and O. Rudolph. The uniqueness theorem for entanglement measures. *J. Math. Phys.*, 43:4252, 2002. 126

[20] P. W. Shor. Equivalence of additivity questions in quantum information theory. *Comm. Math. Phys.*, 246:453, 2003. 126

[21] M. B. Hastings. Superadditivity of communication capacity using entangled inputs. *Nature Physics*, 5:255, 2009. 126

[22] K. G. H. Vollbrecht and R. F. Werner. Entanglement measures under symmetry. *Phys. Rev. A*, 64:062307, 2001. 126

126