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

Last updated: 2019/09/30, 14:34:22
## Contents

0.1 Introductory remarks .................................................. 3  
0.2 Conventions ............................................................ 6  
0.3 Lightning quantum mechanics reminder .......................... 7  

1 Hilbert space is a myth .................................................. 10  
1.1 Mean field theory is product states .............................. 13  
1.2 The local density matrix is our friend ......................... 15  
1.3 Complexity and the convenient illusion of Hilbert space .... 18  

2 Quantifying information and ignorance .......................... 25  
2.1 Relative entropy ....................................................... 32  
2.2 Data compression ..................................................... 36  
2.3 Noisy channels ........................................................ 42  
2.4 Error-correcting codes .............................................. 48  

3 Information is physical ................................................ 52  
3.1 Cost of erasure ........................................................ 52  
3.2 Second Laws of Thermodynamics ................................. 58  

4 Quantifying quantum information and quantum ignorance .... 63  
4.1 von Neumann entropy ................................................. 63  
4.2 Quantum relative entropy .......................................... 66  
4.3 Purification, part 1 .................................................... 68  
4.4 Schumacher compression .......................................... 70  
4.5 Quantum channels .................................................. 72  
4.6 Channel duality ....................................................... 78  
4.7 Purification, part 2 .................................................. 83  
4.8 Deep facts ............................................................. 88  

5 Applications of (mostly) SSA to many body physics ............ 96  

6 Entanglement as a resource ........................................... 101  
6.1 When is a mixed state entangled? ............................... 101  
6.2 States related by LOCC ............................................. 102  
6.3 Entanglement distillation, briefly ............................... 107  
6.4 Distance measures .................................................. 110  
6.5 Zoo of measures of mixed-state entanglement ................. 117  

7 Area laws and local tensor network states ....................... 125
7.1 Local tensor network states .............................................. 126
7.2 Mutual information appreciation subsection ...................... 130
7.3 Small incremental entangling by local Hamiltonians ............ 134

8 Quantum error correction and topological order .................. 137
8.1 Toric code .............................................................. 138
8.2 Entanglement, short and long ....................................... 142
8.3 Shor’s code is a toric code ........................................... 144

9 Tangent vectors to an imagined future ............................... 145
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 and learning to
use 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 (particularly
quantum computation) and ideas about phases of matter.

In the next few paragraphs, I’m going to give some examples of what I mean. Don’t
get scared by the undefined terms, we’ll come back to most of them.

One example is the connection between orders (in the sense of labels on phases of
matter) and memory. The quest for a self-correcting quantum memory (a quantum
hard drive that you can put in your closet without keeping it plugged in), hinges on the
stability of topological order (phases of matter which cannot be distinguished locally)
at finite temperature. More prosaically, the magnetic hard drives we all use as digital
memory rely on spontaneous symmetry breaking.

Another example is the deep antipathy between tractability and ergodicity. Compu-
tationally hard problems (and in particular attempts to solve them with a quantum
adiabatic algorithm), are related to phenomena associated with the word glass. And
integrability, or more generally our ability to solve a model, and hence compute using
classical resources, is, in general, in tension with its ergodicity, i.e. the applicability of
statistical mechanics.
Actually the concept of topological order (in the sense of local indistinguishability of states) is relevant to both the question of applicability of statistical mechanics through the eigenstate thermalization hypothesis, and the notion of quantum error correction.

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. If you are worried about your level of quantum mechanics preparation, do Problem Set 0.5.

**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 (classical Shannon theory).

3. Memory, erasure and the physicality of information.

4. Quantum Shannon theory, distinguishing quantum states (distance measures).

5. Groundstate entanglement area law. Other consequences of locality.
6. Quantum error correction and topological order.

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*, S. Barnett.

Renner and Christandl, notes.

*Quantum channels guided tour*, M. Wolf.


*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

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

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

(rather than \( \sigma_x, y, z \)).

\( \equiv \) means ‘equals by definition’. \( A \overset{1}{=} B \) means we are demanding that \( A = B \). \( A \overset{2}{=} B \) means \( A \) probably doesn’t equal \( B \).

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.

A useful generalization of the shorthand \( \hbar \equiv \frac{\hbar}{2\pi} \) is

\[ dk \equiv \frac{dk}{2\pi}. \]

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

I will try to be consistent about writing Fourier transforms as

\[ \int \frac{d^dk}{(2\pi)^d} e^{ikx} \tilde{f}(k) \equiv \int d^dk \ e^{ikx} \tilde{f}(k) \equiv f(x). \]

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 \( \hbar \) and \( k_B \) are 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 $|\psi\rangle$ 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 linear operators on $\mathcal{H}$ satisfying $A = A^\dagger$ (the adjoint can be defined by $\langle a | A | b \rangle^* = \langle b | A^\dagger | a \rangle$). 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 $\mathbf{H}$:

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

4. When we measure the observable $A$ in the state $|\psi\rangle$, the outcome is an eigenvalue of $A$ ($A |a\rangle = a |a\rangle$), 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 $\text{dim } \mathcal{H} = 2$, 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)):

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

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

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

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, $\| \psi \| = 1$.  

8
$\mathbb{C P}^1$, 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:

$$Z |0\rangle = |0\rangle, \quad Z |1\rangle = -|1\rangle$$

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

$$\left( \begin{array}{c|c} \langle 0 | Z | 0 \rangle & \langle 0 | Z | 1 \rangle \\ \hline \langle 1 | Z | 0 \rangle & \langle 1 | Z | 1 \rangle \end{array} \right) = \left( \begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array} \right) = \sigma^z,$$

the Pauli matrix along $z$. Let’s also define an operator $X$ which satisfies

$$XZ = -ZX \quad \text{(0.2)}$$

(and we’ll demand it too is hermitian and squares to one). Then, given an eigenvector of $Z$,

$$Z |s\rangle = (-1)^s |s\rangle, \quad s = 0, 1, \quad \text{(0.3)}$$

consider the state $X |s\rangle$. It has

$$Z (X |s\rangle) \equiv XZ |s\rangle \equiv -(-1)^s (X |s\rangle).$$

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

$$\left( \begin{array}{c|c} \langle 0 | X | 0 \rangle & \langle 0 | X | 1 \rangle \\ \hline \langle 1 | X | 0 \rangle & \langle 1 | X | 1 \rangle \end{array} \right) = \left( \begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right) = \sigma^x,$$

the Pauli matrix along $x$. We can also define $Y \equiv iXZ$, whose matrix elements are $\sigma^y$. Often it is convenient to conflate the matrix elements $\sigma^{x,y,z}$ and the operators $X, Y, Z$, 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

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

with $a_\mu$ real. We can prove this simply by checking that the final expression parametrizes an arbitrary $2 \times 2$ hermitian 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? (We’ll need to do this for example if we want to let the two systems 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 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’ll get if we assume the state of the combined system is of the form
\[
|a\rangle_1 \otimes |b\rangle_2, \quad (0.4)
\]
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:
\[
(\langle c|_1 \otimes \langle d|_2) (|a\rangle_1 \otimes |b\rangle_2) \equiv \langle c|a\rangle_1 \langle d|b\rangle_2
\]
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 (by definition)
\[
\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2 \equiv \text{span}_\mathbb{C}\{ |a\rangle \otimes |b\rangle, \; |a\rangle_1 \in \mathcal{H}_1, |b\rangle_2 \in \mathcal{H}_2\},
\]
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
\[
|w\rangle = \sum_{a,b} w_{ab} |ab\rangle \neq |v^1\rangle_N \otimes |v^2\rangle_M
\]
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 \(w_{ab} = v^1_a v^2_b\) has rank one. Such a state in \(\mathcal{H}_N \otimes \mathcal{H}_M\) is called unentangled. In all the other states, the two substates are entangled. (An important question for us will be: how entangled?)

\[\begin{align*}
2 & \text{Some folks regard this as a fifth axiom.} \\
3 & \text{This is to be distinguished from the direct sum of two spaces,} \\
\end{align*}\]
\[
\mathcal{H}_N \oplus \mathcal{H}_M \equiv \text{span}_\mathbb{C}\{ |a\rangle_N , |b\rangle_M \}
\]
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

Before saying more about quantum information theory, I want to introduce the kinds of applications to quantum many-body physics I have in mind, and why such ideas are badly needed. After this motivational chapter the discussion will become again elementary, in the sense of starting from nothing, so bear with me.

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 $(L/a)^d$ patches, where $a$ is the size of the patches. So the thermodynamic limit is $L \rightarrow \infty$, 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.\footnote{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 so far is likely to represent a very small sample of the possible emergent quantum systems.}

\footnote{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.}
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.

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 $\beta$, 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, $H \mapsto H + g$? 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.

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 $\text{SU}(2)$ 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 $\partial^2_h f|_{h=0}$, 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.
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.

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

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

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

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

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

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

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

$$|\psi\rangle = \sum_{\{s_x=1,..,D\}} c_{s_1..s_D N} |s_1..s_D N\rangle . \quad (1.1)$$

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

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

$$|\psi_{MF}\rangle = \bigotimes_x \left( \sum_{c_{sx}=1}^{D} c_{sx} |s_x\rangle \right).$$

States which can be factorized in this way (in some factorization of $H$) are called unentangled (with respect to that factorization of $H$). This writes the state in terms of only $N_D$ numbers $c_{sx}$, 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 through a single field (the eponymous 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 $N_D$ 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

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

Here $\langle ij \rangle$ indicates the the site $i$ and $j$ share a link. The first term is a ferromagnetic (if
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:

$$|⇑⟩ ≡ |↑↑↑...⟩, \quad |⇓⟩ ≡ |↓↓↓...⟩$$

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

$$|±⟩ ≡ \frac{1}{\sqrt{2}} (|⇑⟩ ± |⇓⟩)$$

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:

$$|⇒⟩ = \otimes_x |→⟩_x = \otimes_x \left( \frac{|⇑⟩ + |⇓⟩}{\sqrt{2}} \right).$$

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

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 my 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 $|\psi\rangle \neq |a\rangle_A \otimes |b\rangle_B$.

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

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

for any $v^{A,B}$. This is only possible if the coefficient matrix factorizes as $w_{i,m} = v_i^A v_m^B$. 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 $M = M_A \otimes 1_{\bar{A}} = B$ – 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 $\rho_A \equiv \text{tr}_{\bar{A}} |w\rangle \langle w|$. This operation by which we obtained $\rho_A$ is called partial trace.

The density matrix $\rho_A$ is a positive (and hence Hermitian) operator with unit trace. These are general conditions on any density matrix which allow for a probability

8Explicitly,

$$\langle M_A \rangle = \langle w| M_A \otimes 1_{\bar{A}} |w\rangle = \sum_{j,s} \sum_{i,r} w_{jr}^* \langle j|_A \otimes \langle s|_B (M_A \otimes 1_B) w_{ir} |i\rangle_A \otimes |r\rangle_B$$

$$= \sum_{ij,r} w_{ir} w_{jr}^* \langle j|_A M_A |i\rangle_A = \text{tr}_A \rho_A M_A,$$

with

$$\rho_A = \text{tr}_{\bar{A}} |w\rangle \langle w| = \sum_{ij,r} |i\rangle_A \langle j| w_{ir} w_{jr}^* , \quad (\rho_A)_{ij} = \sum_r w_{ir} w_{jr}^*.$$  \hspace{1cm} (1.5)

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

9A positive operator $A$ is one for which $\langle b| A |b\rangle \geq 0$ for all states $|b\rangle$. Beware that one may
interpretation of expectation values $\langle M_A \rangle = \text{tr}_A \rho_A M_A$, and here they follow from the normalizedness of the state $|w\rangle$. As with any hermitian matrix, $\rho_A$ can be diagonalized and has a spectral decomposition:

$$\rho_A = \sum_{\alpha} p_\alpha |\alpha\rangle \langle \alpha|$$

with $\text{tr}_A \rho_A = \sum_{\alpha} p_\alpha = 1$. $p_\alpha \in [0,1]$ 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 $\rho_A$:

$$S[\rho_A] \equiv -\text{tr}_A \rho_A \log \rho_A.$$

---

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 $\mathcal{D}^{|A|}$ 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 encounter an alternative definition that all the singular values ($s$ such that $\det(sI - A) = 0$) 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.
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\(^\text{10}\) it. Some 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, \(f : \{0, 1\}^n \rightarrow \{0, 1\}\). 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 \(\text{dim} \mathcal{H}\).

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

\[^{10}\text{Do you want me to say more about this? Here’s the short version:}\]

\[
\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right) \otimes |0\rangle \xrightarrow{\text{wait}} \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)
\]

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

\[
\frac{1}{\sqrt{2}} \left( |\text{Cat}\rangle + |\text{Dog}\rangle \right) \otimes \frac{|\text{Cat}\rangle + |\text{Dog}\rangle}{\sqrt{2}} \otimes \frac{|\text{Cat}\rangle + |\text{Dog}\rangle}{\sqrt{2}} \otimes \frac{|\text{Cat}\rangle + |\text{Dog}\rangle}{\sqrt{2}}
\]

(You should recognize the observer depicted here from xkcd. The cat pictures are of unknown provenance.)
two) out of $n - 1$ successive pairwise XORs:

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

\[(n + T)^{kT} \geq 2^{2^n}\]  \hspace{1cm} (1.6)

to compute all the functions, so we require

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

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 \( f(n) = e^a n \) 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

\[ \mathcal{H} = \bigotimes_{i=1}^{n} \mathcal{H}_i \]

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:

\[ \mathbf{i} \partial_t |\psi(t)\rangle = H(t) |\psi(t)\rangle \quad (1.7) \]

(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

\[ \mathbf{i} \partial_t U(t) = H(t) U(t), \quad (1.8) \]

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

\[ U(t) = \mathbb{I} - \mathbf{i} \int_0^t dt_1 H(t_1) U(t_1). \]
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:

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

Perhaps we should keep doing this a few more times:

$$U(t) = \mathbb{1} - i \int_0^t dt_1 H(t_1) + (-i)^2 \int_0^t dt_1 \int_0^{t_1} dt_2 H(t_1) H(t_2) + \cdots$$

$$= T \left( \mathbb{1} - i \int_0^t dt_1 H(t_1) + \frac{1}{2}(-i)^2 i \int_0^t dt_1 \int_0^t dt_2 H(t_1) H(t_2) \right)$$

$$+ \frac{1}{3}(-i)^3 i \int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 H(t_1) H(t_2) H(t_3) + \cdots$$

$$= T \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left(-i \int_0^t dt'H(t')\right)^\ell = Te^{-i \int_0^t dt' H(t')}.$$

The tricky step here is the introduction of the *time-ordering* operation $T$: $H(t)$ 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 $T$ puts them in the right order:

$$TH(t_1)H(t_2) = \theta(t_1 - t_2)H(t_1)H(t_2) + \theta(t_2 - t_1)H(t_2)H(t_1)$$

where $\theta(t)$ is the Heaviside function). In that same step, we extend the range of integration from the simplex $t_1 \geq t_2 \geq t_3 \geq \cdots \geq t_\ell \geq 0$ 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 $k$ qubits at a time. The time evolution operator from time 0 to time $t$ is

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

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 $\Delta t$, which is fine if $\Delta t \ll \| \partial_t H \|^{-1}$. Here $\| O \| \equiv \sup_{\{\text{normalized states}, \psi\}} \| O \psi \|$ 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

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

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 $\epsilon$ 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 $n$ qubits. 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 $n^\alpha$ for some $\alpha$. The number of circuits we can make from these is (just as in the classical case (1.6))

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

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, $N_{\text{circuits}}$ bounds from above the number of distinct states we can make.
Let’s allow an error $\epsilon$, so we declare victory if we get inside a ball of radius $\epsilon$ 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

$$V_\epsilon = \epsilon^D \frac{\pi^{D/2}}{\Gamma(D/2 + 1/2)} \approx \epsilon^{2^2n} \frac{\pi^{2n}}{\Gamma(2^n)}.$$

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

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

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

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

a doubly-exponentially tiny fraction. It’s the powers of $\epsilon$ 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 $\mathcal{H} = \bigotimes_{i=1}^N \mathcal{H}_x$ 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

$$\mathcal{H} = \text{span}\{|s_1s_2s_3\cdots\}, \ s_i = 0 \text{ or } 1$$

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.11 (Propose a countable basis. Then line up the basis elements in a big vertical table. Make a new number by flipping the $n$th digit of the $n$th 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.

11 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.

11 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 and ignorance

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).
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 $P(n, l) = \text{Prob}(A_n, B_l)$. The distribution for the second variable (ignoring the first) is

\[q_l \equiv \sum_n P(n, l).
\]

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

\[p(n|l) \equiv \frac{P(n, l)}{q_l}.
\]

(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:

$$H(X,Y) = H(Y) + H(X|Y).$$

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:

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

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.:

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

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 ↑ or ↓, 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 ↑ if $A = \uparrow$: $p(D_\uparrow|A_\uparrow) = 1$ but if $A = \downarrow$, it says ↓ only $3/4$ of the time: $p(D_\downarrow|A_\downarrow) = 3/4$. The question is: if the detector says ↑, what probability should we assign to the statement that $A$ is actually ↑?

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 ↑ both times. Bayes rule generalizes to

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

and we get a more certain outcome:

$$p(A_\uparrow|D_\uparrow^1D_\uparrow^2) \propto \frac{p(D_\uparrow^1D_\uparrow^2|A_\uparrow)p(A_\uparrow)}{p(D_\uparrow^1|A_\uparrow)p(D_\uparrow^2|A_\uparrow)} = 1 \cdot 1 \cdot \frac{1}{4}$$
\[ p(A_\uparrow | D_1 \uparrow D_2 \uparrow) \propto p(D_1 \uparrow D_2 \uparrow | A_\uparrow) p(A_\uparrow) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \]

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

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

\[
\begin{align*}
H(1/4, 3/4) &= .56 \\
H(4/7, 3/7) &= .68 \\
H(16/19, 3/19) &= .44 \\
H(64/67, 3/67) &= .18
\end{align*}
\]

... \( H\left(\frac{4^n}{3 + 4^n}, \frac{3}{3 + 4^n}\right) \) \( \rightarrow \infty \).

Exercise: How does \( H(n) = H\left(\frac{4^n}{3 + 4^n}, \frac{3}{3 + 4^n}\right) \) decay as \( n \rightarrow \infty \)? 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 \( \uparrow \) and not \( \downarrow \). For better or worse, fire-breathing is recessive, so an arctopus with phenotype \( \uparrow \) necessarily has genotype \( \uparrow\uparrow \), while a non-fire-breathing arctopus may be \( \downarrow\uparrow \), \( \uparrow\downarrow \) or \( \downarrow\downarrow \).

If we breed a fire-breathing 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 \( \downarrow\uparrow \) or \( \uparrow\downarrow \). If the offspring does not breathe fire then maybe the father was \( \downarrow\downarrow \). 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 \( \uparrow\downarrow \) and \( \downarrow\uparrow \) in the labelling.)

Now, if we repeatedly mate these arctopuses, we have

\[ p(\text{ith offspring does not breathe fire}| \text{dad is } \downarrow\downarrow) = 1 \]

\[ p(\text{ith offspring does not breathe fire}| \text{dad is } \uparrow\downarrow) = 1/2. \]
If, as is likely, the first offspring does not breathe fire (I’ll write this as $x_1 = \downarrow$), we infer

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

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

which when we normalize gives

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

If the second offspring also comes out $\downarrow$, we update again:

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

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

so now we assign $p(\downarrow \downarrow \mid ...) = 2/3$. 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...
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 ↓, the deity makes one sealed room containing an intelligent person with red hair. If the coin says ↑ 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 ↑, \( p(\downarrow | G) = 0 \). The problem is: if your hair is red, what probability should you assign to ↑, 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

\[
p(\uparrow | R) \geq \frac{1}{101}.
\]

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 \( p(R | \uparrow) = 1, p(R | \downarrow) = \frac{1}{100} \))

\[
p(\uparrow | R) = \frac{1}{100} \cdot \frac{101}{101} = \frac{1}{2}.
\]

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 $H[p] = \int dx (-p(x) \log p(x))$.

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 $\Delta x$. But the Shannon entropies of those distributions actually approach

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

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 $\Delta x$ (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 ...
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 $p \leftrightarrow q$.  

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

$$-D(p||q) = \sum_{x \in \mathcal{X}} p_x \log \frac{q_x}{p_x} = \sum_{x \in A} p_x \log \frac{q_x}{p_x}$$

$$\leq \sum_{x \in A} p_x \left( \frac{q_x}{p_x} - 1 \right) = \sum_{x \in A} (q_x - p_x) = \sum_{x \in A} q_x - 1 \leq 0.$$ 

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)):

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

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,

$$I(X : Y) = \sum_{xy} p_{xy} \log \frac{p_{xy}}{p_xp_y} = \left\langle \log \left( \frac{p(X,Y)}{p(X)p(Y)} \right) \right\rangle_{XY}$$

$$= -\sum_{xy} p_{xy} \log p_x + \sum_{xy} p_{xy} \log p(x|y) = H(X) - H(X|Y). \quad (2.5)$$

\(^{12}\)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

\[
I(X : X) = H(X) - H(X|X) = H(X)
\]

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

\[
I(X : Y) \geq 0
\]

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):

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

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 \( p_{yx} = \binom{0}{a} \binom{b}{b} \), 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”:

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

which is saturated by the completely uncorrelated distribution \( p_{x_1 \cdots x_n} = p_{x_1} \cdots p_{x_n} \).

This is sometimes also called subadditivity of the entropy.

Here is a useful mnemonic\(^{13} \):

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 \( O_{X,Y} \) of random variables \( X,Y \),

\[
I(X : Y) \geq \frac{1}{2} \langle O_X O_Y \rangle_c^2 \leq \frac{1}{2} \| O_X \|^2 \| O_Y \|^2.
\]

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\(^{14} \)

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

Later in \S 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.

---

\(^{13}\) There are some shortcomings of using a Venn diagram to illustrate entropies. I’ll explain below in \S 2.3.1

\(^{14}\) 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 \( H(p) = -\sum p \log p \) 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 \( -\log p \) 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.

<table>
<thead>
<tr>
<th>Move number</th>
<th>outcome</th>
<th>( p_{hit} )</th>
<th>( p_{miss} )</th>
<th>info gained = (-\log p_{outcome})</th>
<th>total info gained so far</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>miss</td>
<td>( \frac{1}{64} )</td>
<td>( \frac{63}{64} )</td>
<td>(-\log \frac{63}{64})</td>
<td>( \log \frac{64}{63} )</td>
</tr>
<tr>
<td>2</td>
<td>miss</td>
<td>( \frac{1}{64} )</td>
<td>( \frac{62}{64} )</td>
<td>(-\log \frac{62}{64})</td>
<td>( \log \frac{64}{63} )</td>
</tr>
<tr>
<td>3</td>
<td>miss</td>
<td>( \frac{1}{64} )</td>
<td>( \frac{62}{64} )</td>
<td>(-\log \frac{62}{64})</td>
<td>( \log \frac{64}{63} )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>32</td>
<td>miss</td>
<td>( \frac{1}{33} )</td>
<td>( \frac{32}{33} )</td>
<td>(-\log \frac{32}{33})</td>
<td>( \log \frac{64}{63} )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>48</td>
<td>miss</td>
<td>( \frac{1}{17} )</td>
<td>( \frac{16}{17} )</td>
<td>(-\log \frac{16}{17})</td>
<td>( \log \frac{64}{63} )</td>
</tr>
<tr>
<td>49</td>
<td>hit</td>
<td>( \frac{1}{16} )</td>
<td>( \frac{15}{16} )</td>
<td>(-\log \frac{1}{16})</td>
<td>( 2 + \log \frac{16}{15} = 6 )</td>
</tr>
</tbody>
</table>

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 \( \alpha \) is a value of the RV \( A \), then this \( x \) is an RV in \( A^N \).) The number of appearances of a particular \( \alpha \) is about \( Np_\alpha \). At
large $N$ we can ignore fluctuations about this average, and ignore the fact that $Np_\alpha$
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

$$\frac{1}{N} \log \Omega \underset{N \gg 1}{\approx} \frac{1}{N} \left( N \log N - \sum_\alpha (Np_\alpha) \log (Np_\alpha) \right)$$

$$= - \sum_\alpha p_\alpha \log p_\alpha = H(p). \quad (2.6)$$

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

\[ P(n) = \binom{N}{n} p^n (1-p)^{N-n} \quad \text{for} \quad n \geq 1,\quad e^{-\frac{(n-\langle n \rangle)^2}{2\sigma^2}}, \quad \langle n \rangle = Np, \sigma = \sqrt{Np(1-p)} \]

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 \( S_\delta \) by throwing in the elements of \( \mathcal{X} \) starting from the most probable, until the total probability missing is \( \delta \). That is: \( S_\delta \) 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 \( S_\delta \). 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 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 \( p_\alpha \) and doesn’t tell us which \( \alpha \) 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 \( \alpha \) it is. The answers to the questions then comprise the binary digits of an efficient binary code for each element \( \alpha \) in the sample set \( \{\alpha\} \). Efficiency means minimizing the average code length

\[ \langle \ell \rangle \equiv \sum_\alpha p_\alpha \ell_\alpha \]
where \( \ell_\alpha \) is the number of questions needed to identify uniquely element \( \alpha \).

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>
<thead>
<tr>
<th>( x )</th>
<th>( p_x )</th>
<th>dumb code</th>
<th>Shannon optimal code</th>
<th>( - \log p_x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( \frac{1}{2} )</td>
<td>00</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>( \frac{1}{4} )</td>
<td>01</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>( \frac{1}{8} )</td>
<td>10</td>
<td>110</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>( \frac{1}{8} )</td>
<td>11</td>
<td>111</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 1**

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 \( \ell_{\text{max}} \) questions, thereby removing all the potential daughter leaves. A codeword of length \( \ell \) eliminates \( D^{\ell_{\text{max}}-\ell} \) terminating daughter leaves (at depth \( \ell_{\text{max}} \)). The number of terminating leaves of the tree which are not codewords is then

\[
\sum_\alpha D^{\ell_{\text{max}}-\ell_\alpha} \leq D^{\ell_{\text{max}}}
\]

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

\[ \sum_{\alpha} D^{-\ell_{\alpha}} \leq 1. \]  

(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 \textit{uniquely decodable}. (That is: a code \( X \) is uniquely decodable if \( X^N \) is \textit{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 inequality holds for prefix codes).

Here’s why [C&T p.116-117]: Consider

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

and gather the terms by total length, \( m \):

\[ = \sum_{m=1}^{k\ell_{\max}} a(m) D^{-m} \leq k\ell_{\max}. \]

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 \( \forall k \),

\[ \sum_{x \in X} D^{-\ell_x} \leq (k\ell_{\max})^{1/k} \xrightarrow{k \to \infty} 1. \]

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, \textit{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 \( \ell_{\alpha} \) are integers and assuming (2.7) is saturated, imposing it with a Lagrange multiplier \( \lambda \):

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

is extremized when

\[ 0 = \partial_{\ell_{\alpha}} J|_{\ell=\ell^*} = p_{\alpha} - \lambda \log DD^{-\ell^*_{\alpha}} \implies D^{-\ell^*_{\alpha}} = \frac{p_{\alpha}}{\lambda \log D} \]
but the constraint determines $1 = \sum_{\alpha} D^{-\ell_{\alpha}} = \frac{1}{\lambda \log D} \sum \rho_{\alpha} = \frac{1}{\lambda \log D}$ so we get $\ell_{\alpha}^* = -\log_{D} \rho_{\alpha}$ and

$$\langle \ell \rangle^{*} = \sum_{\alpha} \rho_{\alpha} \ell_{\alpha}^{*} = - \sum_{\alpha} \rho_{\alpha} \log_{D} \rho_{\alpha} = H_{D}(p).$$

And the extremum is actually a minimum: $\langle \ell \rangle \geq H_{D}[p]$. 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

$$\langle \ell \rangle - H_{D}(p) = \sum_{\alpha} \rho_{\alpha} \frac{\ell_{\alpha}}{\log_{D}(D^\ell_{\alpha})} + \sum_{\alpha} \rho_{\alpha} \log_{D} \rho_{\alpha}$$

$$= \sum_{\alpha} \rho_{\alpha} \log_{D} \left( \frac{\rho_{\alpha}}{q_{\alpha}} \right) + - \log_{D} \left[ \sum_{\alpha} D^{-\ell_{\alpha}} \right]$$

$$\equiv D(p||q) \geq 0$$

(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: \(^{15}\)

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p_{x}$</th>
<th>.25</th>
<th>.3</th>
<th>.45</th>
<th>.55</th>
<th>codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_{x}$</td>
<td>1</td>
<td>01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.25</td>
<td>.25</td>
<td>.3</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.2</td>
<td>.25</td>
<td>.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.15</td>
<td>.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{15}\)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 \(i\)th interval, we need to keep \(\lceil \log \frac{1}{p_i} \rceil\) 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 \(\langle l \rangle = H(X)\). 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

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

(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 $I(A : B|E) = H(A|E) - H(A|BE)$. \(^{16}\)

\(^{16}\)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 $H(A|(BE))$.  

44
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

\[
H(ABE) = H(A) + H(E|A) + H(B|A,E)
\]

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 that it is still possible to read this sentence though all the vowels have been removed. 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

$$\log N_E \overset{\text{Stirling}}{\approx} N_0H(q)$$

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

$$N = \sum_{k=0}^{\infty} N_0H(q)^k = \frac{N_0}{1 - H(q)}$$

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

$$2^{N_0} = 2^{N(1-H(q))}$$

distinct messages.

Each transmitted bit carries only

$$\frac{1}{N} \log (2^{N(1-H(q))}) = 1 - H(q)$$

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.

**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 number of reliably sendable messages is

\[\begin{align*}
\text{# of reliably sendable messages} &= 2^{N(H(B)−H(B|A))} = 2^{N(H(A)−H(A|B))} = 2^{NI(A:B)}.
\end{align*}\]

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 $a_i^j$,

$$
\begin{pmatrix}
  a_1^1 & \cdots & a_1^m \\
  a_2^1 & \cdots & a_2^m \\
  \vdots & \ddots & \vdots \\
  a_n^1 & \cdots & a_n^m \\
\end{pmatrix}
\begin{pmatrix}
  (\sum_j a_i^j)_2 \\
  (\sum_j a_i^j)_2 \\
  \vdots \\
  (\sum_j a_i^j)_2 \\
\end{pmatrix}
$$

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

$$
\begin{pmatrix}
  s_1 \\
  s_2 \\
  s_3 \\
  s_4 \\
\end{pmatrix}
$$

send

$$
\begin{pmatrix}
  1 \\
  1 \\
  1 \\
  1 \\
\end{pmatrix}
\begin{pmatrix}
  s_1 \\
  1 \\
  1 \\
  s_2 \\
\end{pmatrix}
\equiv
\begin{pmatrix}
  1 \\
  1 \\
  1 \\
  0 \\
\end{pmatrix}
= Gs
$$

(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 

$$H \equiv \begin{pmatrix} P \\ \mathbb{I}_{4 \times 3} \end{pmatrix}.$$ 

(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

$$H = -Z_1Z_2Z_3Z_5 - Z_1Z_2Z_4Z_6 - Z_2Z_3Z_4Z_7.$$ 

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:

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

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. 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

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

---

17For 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
\[
\text{Xerox} |b\rangle \otimes |\text{anything}\rangle = |b\rangle \otimes |b\rangle.
\]

But then what does it do to the superposition? By assumption, it copies it:
\[
\text{Xerox} \left( \frac{|a\rangle + |b\rangle}{\sqrt{2}} \otimes |\text{anything}\rangle \right) = \left( \frac{|a\rangle + |b\rangle}{\sqrt{2}} \right) \otimes \left( \frac{|a\rangle + |b\rangle}{\sqrt{2}} \right).
\]

But that’s not the same as the superposition of the images:
\[
\text{Xerox} \left( \frac{|a\rangle + |b\rangle}{\sqrt{2}} \otimes |x\rangle \right) \neq \frac{1}{\sqrt{2}} \left( |a\rangle \otimes |a\rangle + |b\rangle \otimes |b\rangle \right)
= \frac{1}{\sqrt{2}} \left( \text{Xerox} |a\rangle \otimes |x\rangle + \text{Xerox} |b\rangle \otimes |x\rangle \right).
\]

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 \( \S 8 \).