Introduction to Classical and Quantum Information Theory

and other random topics from probability and statistics

Sam Kennerly
4 September 2009
Drexel University PGSA informal talk

0.0 DNA and Beethoven’s 9th Symphony

- In my last presentation, I said the information content of the human genome is about equal to a recording of Beethoven’s 9th.

- **3 billion base pairs** in human DNA, each occupied by **1 of 4 bases**. Representing each base by two binary digits, we need (2 bits)*(3 billion) = **6 gigabits** = 750 MB of disk space to sequence a genome.

- An audio CD records two 16-bit samples every 44,100th of a second. The 9th is about 72 minutes long, so it needs (2)(16)(44,100)(72)(60) bits = **6 Gb**.

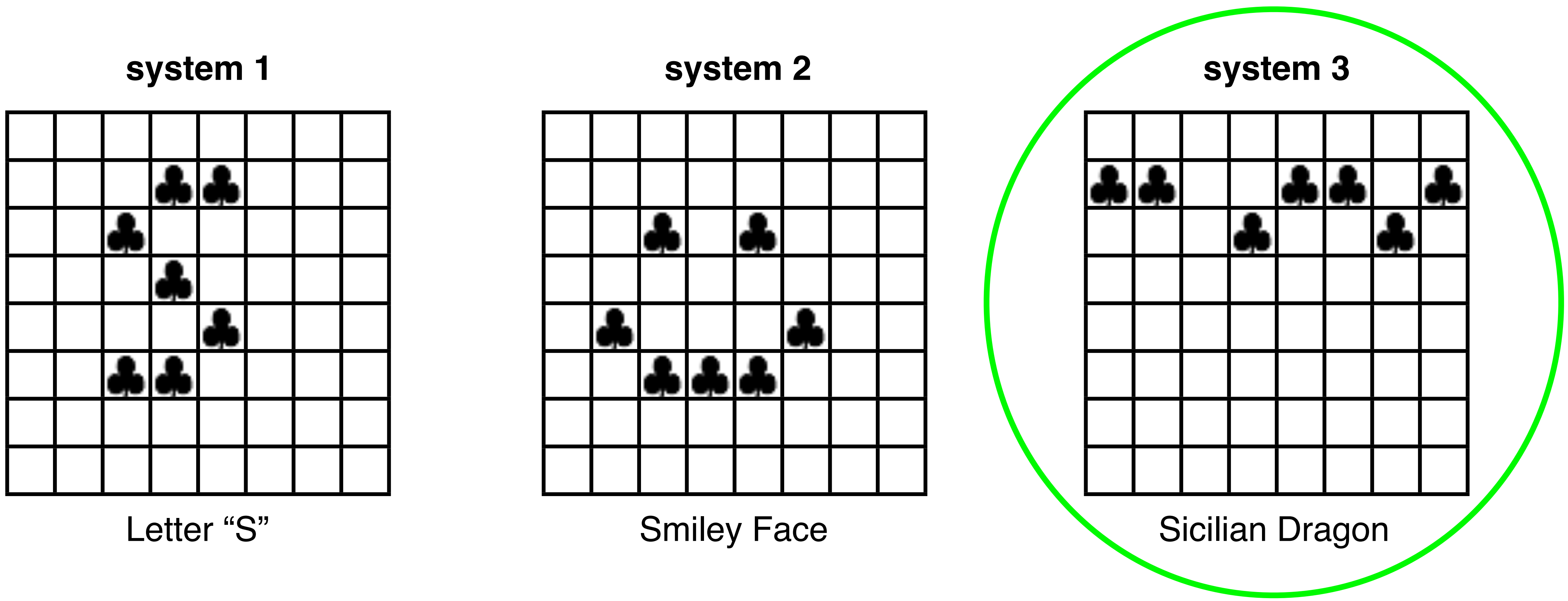
- Question: Do we really need **all** those bits? Can’t we .zip them or something?

0.1 DNA and Beethoven's 9th Symphony

- DNA answer: The *entropy rate* of DNA is about 1.7 bits per base, about 85% of the maximum 2 bits/base. Shannon's **source coding theorem** says that no algorithm can compress the genome to less than (0.85)(750MB) = 637.5 MB.
- Real-life compression is imperfect; source-coding theorem gives a **lower bound** on file size. Compression schemes designed for one type of data may work poorly for others. (ZIP is notoriously bad for audio encoding.)
- Beethoven answer: The entropy rate depends on the recording, but existing *Golumb-Rice* encoders compress to about 50-60% original size.
- **Lossy compression** can make files smaller, but **information is destroyed!** Examples: mp3/aac/ogg (audio), jpg/gif (graphics), DivX/qt/wmv (video)
- Experiments suggest VBR-mp3 at 18% is good enough to trick listeners.
- How much of DNA info is "junk" is debated; 95% is a popular estimate.

1.0 What is entropy?

- Old-fashioned answer: Entropy is a measure of how disordered a system is.
- Dilemma: How do we define disorder? A broken egg is more disordered than a not-broken egg... but which of the following pictures is least disordered?



- Moral of story: **Disorder is in the eye of the beholder.**

2.0 Boltzmann's entropy

- Question: How do we model the behavior of gases in a steam engine?

- 1 L of ideal gas at STP has 2.7 10^{22} molecules. If each has 3 position and 3 momentum coordinates, differential eqn. of motion has ~ 10^{23} variables. (Actual gases are much more complicated, of course.)

- Solving this equation is an impractical way to build locomotives.

- Answer: Call each configuration the system a **microstate.** If two different microstates have the same Energy, Volume, and Number of particles, call them equivalent. A **macrostate** a set of microstates with the same (E,V,N) values.

- M**ultiplicity** Ω(E,V,N) is the number of microstates for a given macrostate.

- Ω is measure of how much information we are ignoring in our model of the system. For this reason, I like to call it the **ignorance** of a macrostate.

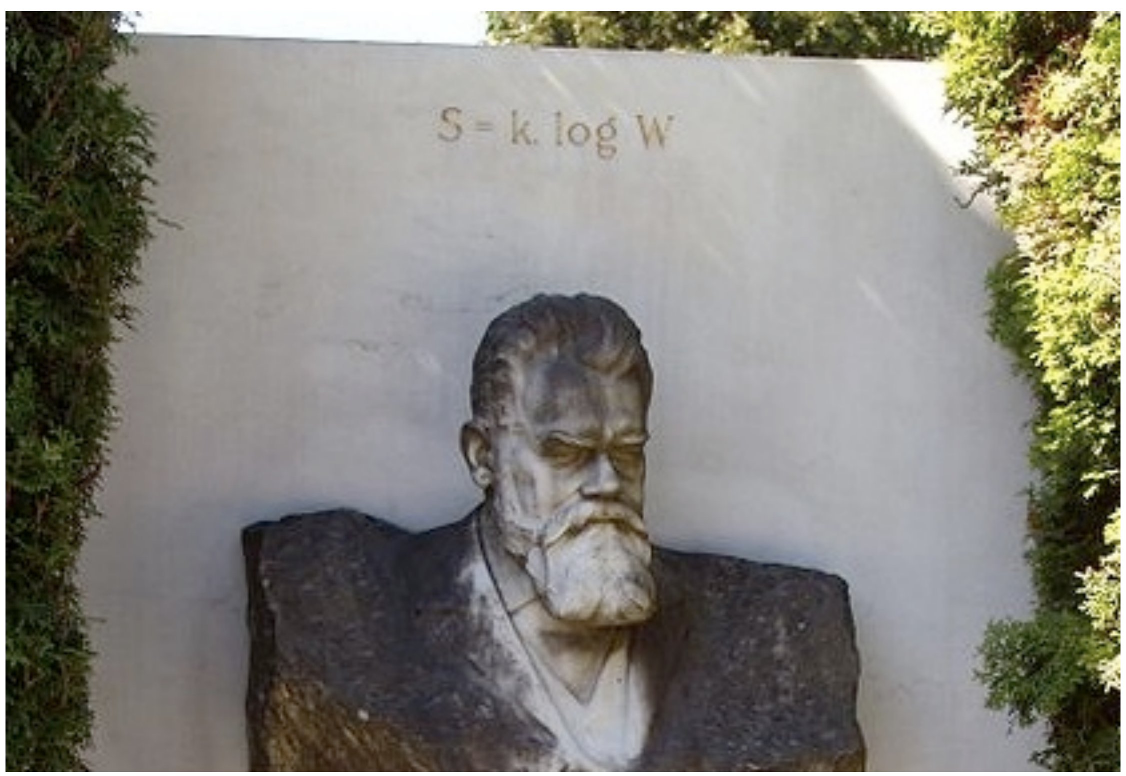
2.1 Boltzmann's entropy

- This method of counting microstates per macrostate is called **microcanonical ensemble theory.** Boltzmann defined the entropy of a macrostate like so:

$$S(E, V, N) = k \ln(\Omega)$$

- This entropy is the **logarithm of ignorance** times a constant $k \approx 1.38\ 10^{23}$ J/K .

- To help us remember this formula, Boltzmann had it carved into his tombstone.



- This is Ludwig Boltzmann's tomb in Vienna.

  (Apparently he was one of those people who prefer "log" to "ln." Also he used W for multiplicity, but you get the idea.)

- Boltzmann's kinetic theory of gases caused some controversy because it apparently requires systems to be inherently discrete.

- Quantum-mechanical systems with discrete energy levels fit nicely into this theory!

3.0 Shannon's entropy

- In 1937, Claude Shannon wrote a famous Master's thesis about using Boolean algebra to write computer programs. During WWII he worked with Alan Turing on cryptography and electronic control theory for Bell Labs.

- Shannon later published his **source-coding** and **noisy-channel** theorems. These placed limits on file compression and the data capacity of a medium subject to noise and errors. Both theorems use this definition of entropy:

$$S[p_n] = -\sum_n p_n \log(p_n)$$

for discrete probability distributions

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

for continuous probability distributions

- **Gibbs' entropy** from thermodynamics is Shannon's entropy times *k*,* though Shannon's entropy is defined for probability distributions, not physical states. *S* is a measure of **how much information is revealed by a random event**.

* Prof. Goldberg and I opine that temperatures should be written in Joules, in which case *k = 1.*

3.1 Shannon's entropy

- For a random variable *X,* a continuous probability distribution *p(x)* is defined:

$$P[a \leq X \leq b] = \int_a^b p(x)\ dx$$

- A probability distribution *p(x)* is also called a **probability density function** or **PDF**. (Technically *p(x)* doesn't have to be a function as long as it can be integrated. For example, Dirac's *δ(x)* is a valid PDF but not a function.)
- From the definition it follows that $p(x) \geq 0$ and $\int_{-\infty}^{+\infty} p(x)\ dx = 1$.
- Example: Cryptographers perform **frequency analysis** on ciphertexts by writing a discrete PDF for how often each letter appears. For a plaintext, this PDF has non-maximal entropy; the letter "E" is more probable than "Q."
- Example: Password entropy is maximized by using uniformly-chosen random letters instead of English words. Including numbers and symbols increases *S*.

3.2 Shannon's entropy

- To better understand Shannon's entropy, first define a **surprisal** $I_n = \log(p_n^{-1})$ for each possible random outcome p_n .

- Example: Alice rolls two dice at the same time. Bob bets her $1 that she will not roll "boxcars" (two 6's). If Alice wins, Bob's surprisal will be log(36).

- Example: The table below shows how surprised we should be when dealt certain types of Texas Hold 'Em hands preflop.

hand	AA	AA/KK	99 or better	any pair	any suited	the hammer
surprisal	log(221)	log(111)	log(37)	log(17)	log(4.25)	log(111)

- Shannon's entropy for a PDF is **the expectation value of surprisal**.

$$\left\langle \log\left(\frac{1}{p_n}\right)\right\rangle = -\left\langle \log(p_n)\right\rangle = -\sum_n p_n \log(p_n)$$

IMPORTANT TECHNICALITY: **0 log(0) = 0**. Use l'Hôpital's rule and $\lim_{x\to 0}[x\log(x)] = \lim_{y\to\infty}[\log(y)/y]$.

3.3 Shannon's entropy

◇ Question: What base to use for log?

◇ Answer: Any number! Information entropy comes in dimensionless units.

base	2	*e*	10
unit name	**bit**	**nat**	**hartley** (or **ban**)

♣ Shannon is credited with inventing the term "bit" for the entropy of a single fair coin toss.

♣ Ralph Hartley was a Bell Labs information-theorist working with Turing and Shannon.

◇ Question: Why use a logarithm in the definition of entropy?

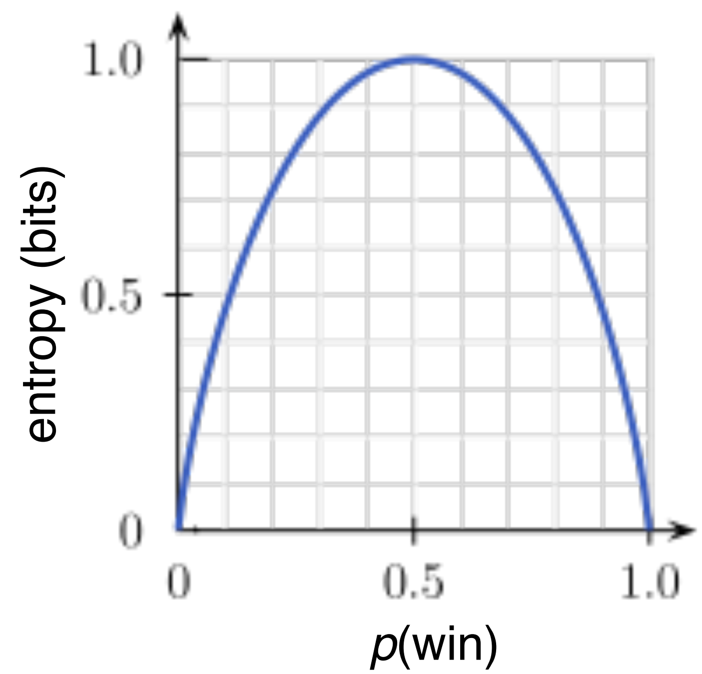
◇ Answer: Observing N outcomes of a random process should give us N times as much information as one outcome. Information is an **extensive quantity**.

◇ Example: Rolling a die once has 6 possible outcomes and rolling it twice has 6^2 outcomes. The entropy of two die rolls is $\log(6) + \log(6) = \log(6^2) \approx 5.17$.

3.4 Shannon's entropy

- The entropy of a fair coin toss is (.5)(log 2) + (.5)(log 2). In base 2, that's **1 bit**.
- The entropy of an *unfair* coin toss is given by the **binary entropy function**.
- 2-player Hold 'Em preflop all-in hands are examples of unfair coin tosses:

hand	p(win)	surprisal	entropy
AA vs AKs	87%	2.9	0.557 bit
AKo vs 89s	59%	1.3	0.976 bit
89s vs 44	52%	1.1	0.999 bit
44 vs AKo	54%	1.1	0.996 bit
KK vs 88	80%	2.3	0.722 bit



Here best **hand** is written first, **p(win)** is prob. best hand wins, and **surprisal** is $\log_2([1\text{-}p(\text{win})]^{-1})$.

3.5 Shannon's entropy

- For an *N*-sided fair die, each outcome has surprisal *N*. The entropy is

$$-\sum_{1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) = \sum_{1}^{N} \frac{1}{N} \log(N) = \log(N)$$

so Boltzmann's entropy is just Shannon's entropy for a uniform discrete PDF.

- If *p(x)* is zero outside a certain range, *S* is maximal for a **uniform distribution.** (Of course! A fair die (or coin) is inherently less predictable than an unfair one.)

- For a given standard deviation σ, *S* is maximal if *p(x)* is a **normal distribution.** In this sense, bell curves are "maximally random" - but be *very careful* interpreting this claim! Some PDFs (e.g. Lorentzians) have no well-defined σ.

- For multivariate PDFs, *Bayes' theorem* is used to define **conditional entropy**:

$$p(x|y) = \frac{p(x)}{p(y)} p(y|x) \quad \Rightarrow \quad S[X|Y] = -\sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(y)}\right)$$

4.1 Thermodynamics

- Recall how temperature is defined in thermodynamics: $\frac{1}{T} \equiv \left(\frac{\partial S}{\partial U} \right)_{N,V}$

- Define **coldness*** β = 1 / T . Given a system with fixed particle number and volume, find the probability of each state as a function of internal energy *U*.

- Find Shannon's entropy for each PDF, then find β = (∂S/∂U). The result is an information-theoretical definition of temperature in Joules per nat!

- In other words, coldness is a measure of **how much entropy a system gains when its energy is increased**. Equivalently, *T* is a measure of how much energy is needed to increase the entropy of a system.

- It is energetically "cheap" to increase the entropy of a cold system. If a hot system gives energy to a cold one, the total entropy of both systems increases. The observation that heat flows from hot things to cold leads to the 2nd Law...

* Coldness is more intuitive when dealing with negative temperatures, which are *hotter* than ∞ Kelvins!

4.2 Thermodynamics

- There have been many attempts to clearly state the 2nd Law of Thermo:
- Statistical: The entropy of a closed* system at thermal equilibrium is more likely to increase than decrease as time passes.
- Clausius: “Heat generally cannot flow spontaneously from a material at lower temperature to a material at higher temperature.”
- Kelvin: “It is impossible to convert heat completely into work in a cyclic process.”
- Murphy: “If there’s more than one way to do a job, and one of those ways will result in disaster, then somebody will do it that way.”
- My attempt: “Any system tends to acquire information from its environment.”

* **Loschmidt’s paradox** points out that if a system is truly “closed,” i.e. it does not interact with its environment in any way, then the statistical version of the 2nd Law violates time-reversal symmetry!

5.0 Von Neumann's entropy

- Despite his knowledge of probability, Von Neumann was reportedly a terrible poker player, so he invented **game theory**.
- Imagine playing 10,000 games of rock-paper-scissors for $1 per game. **Pure strategies** can be exploited: if your opponent throws only scissors, you should throw only rocks, etc. The best option is a **mixed strategy** in which you randomly choose rock, paper, or scissors with equal probability.
- Assume your opponent knows the probability of each of your actions. The entropy of a pure strategy is 0. The entropy of 1/3 rock + 1/3 paper + 1/3 scissors is $\log(3) \approx 1.58$ bits, which is the maximum possible for this game.
- Von Neumann's poker models (and all modern ones) favor mixed strategies. But unlike rock-paper-scissors, the best strategy is *not* the one that maximizes entropy. The best poker players **balance** their strategies by mixing profitable plays with occasional entropy-increasing bluffs and slowplays.

5.1 Von Neumann's entropy

- Von Neumann (and possibly also Felix Bloch and Lev Landau) developed an alternate way to write quantum mechanics in terms of **density operators.**
- Density operators are useful for describing *mixed states* and systems in thermal equilibrium. The related *von Neumann entropy* is also used to describe entanglement in quantum computing research.
- Density operators are defined as real combinations of **projection operators**. A projection P is a linear operator such that $P = P^2$ ($= P^3 = P^4 = \ldots$)
- For any vector Ψ, there is a projection P_Ψ . In Dirac notation, $\hat{P}_\Psi = |\Psi\rangle\langle\Psi|$. This notation says, "Give P_Ψ a vector. It will take the inner product of that vector with Ψ to produce a number, and it will output Ψ times that number."

$$|a\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \hat{P}_a = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad |b\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \Rightarrow \hat{P}_b = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

5.2 Von Neumann's entropy

- A **pure** quantum state can be represented by some state vector Ψ in some complex vector space. Its density operator is defined $\rho = P_\Psi$.

- **Mixed** quantum states represent uncertain preparation procedures. For example, Alice prepares a spin-1/2 particle in the S_z eigenstate $|\uparrow\rangle$. Chuck then performs an S_x measurement but *doesn't* tell Bob the result. Bob knows the state is now either $\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$ or $\frac{1}{\sqrt{2}}(|\uparrow\rangle - |\downarrow\rangle)$, but he doesn't know which!

- Bob can still write a density operator for this mixture of states. He constructs a projection operator for each possible state, then multiplies each operator by 50% and adds the two operators together:

$$\hat{P}_1 = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \hat{P}_2 = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \Rightarrow \quad \hat{\rho} = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- In general, a density operator is defined $(p_1)\hat{P}_1 + (p_2)\hat{P}_2 + (p_3)\hat{P}_3 \cdots$ where each P is the projection of a state and each p the probability the system is in that state.

5.3 Von Neumann's entropy

- If Bob measures the z-spin of his mixed-state particle, the expectation value of his measurement is the *trace* of the operator [S_z][ρ]. (For a matrix, trace is the sum of diagonal elements. In this case, that would be 0.)

- The diagonal elements of ρ are the probability of Bob finding S_z to be +½ or -½. If Bob wants to know the probability of finding the result of some other measurement, he rewrites ρ using the eigenstates of that operator as his basis.

- The time-evolution of ρ follows the **Von Neumann equation**, the density-operator version of the Schrödinger equation: $\imath\hbar\,\partial_t\hat{\rho} = [\hat{H}, \hat{\rho}]$

- **Von Neumann's entropy** is defined by putting ρ into Shannon's entropy:

$$S = -\sum_n p_n \log(p_n) = -Tr[\hat{\rho}\log(\hat{\rho})]$$

- Performing an observation changes ρ in such a way that *S always* increases!

5.4 Von Neumann's entropy

- Question: How do you find the log of an *operator* ?
- Answer: If the operator is Hermitian, it can be diagonalized by a unitary transformation $H = U^{-1}DU$. Since Exp[$U^{-1}DU$] = U^{-1} Exp[D] U , we can "log" an operator by finding the log of its eigenvalues and then similarity transforming.
- A projection P_ψ made from a vector Ψ is always Hermitian. A real combination of Hermitian operators is also Hermitian, so ***ρ*** **is Hermitian**. In fact, all its eigenvalues are in the interval [0,1] (Remember, zero eigenvalues can be ignored in the entropy formula because 0 log(0) = 0.)
- The definition of ρ can be used to prove that its trace Tr[ρ] = 1 always.
- The quantum version of **canonical ensemble** thermodynamics uses density operators. The **partition function** Z and density operator ρ are given by:

$$Z = Tr[EXP(-\beta\hat{H})] \qquad \hat{\rho} = \frac{1}{Z}\, EXP(-\beta\hat{H})$$

6.0 Quantum information paradoxes

- According to the Schrödinger, Heisenberg, and Von Neumann equations, quantum time evolution is unitary. Unitary transformations are *always* invertible, which means they can *never* destroy information about a state.

- The Copenhagen interpretation, however, says that measuring a system "collapses" it into an eigenstate. This time evolution is a projection onto a vector, so it is singular. Singular transformations *always* destroy information. Schrödinger thought this "damned quantum jumping" was absurd.

- Von Neumann's entropy is *increased* by projective measurements. Does this help solve Schrödinger's objection? If entropy is the amount of random information in a system, perhaps measurements only scramble information.

- Hawking, 't Hooft, Susskind, and Bekenstein claim that black holes maximize entropy for a given surface area, and if one of two entangled particles is sucked into the horizon, Hawking radiation is emitted as a *mixed* state. This is not unitary time-evolution either! Do black holes count as observers?

Something Completely Different

- Humans seem to be naturally inept at understanding certain concepts from probability and statistics. Some notorious examples are below:

- 1. Figher pilots at a particular airbase are each shot down with probability 1% on each mission. What are the odds that a pilot completes 200 missions?

- 2. Betting on a number in roulette pays 35:1. There are 38 numbers on an American roulette wheel. What is the expectation value of 100 bets on red 7?

- 3. You are offered 3 doors to choose from on a game show. Behind one is a car; the other two contain goats. Your host, Monty, chose the winning door before the show by throwing a fair 3-sided die. After you choose a door, Monty will open another door. This door will always reveal a goat, and Monty will ask if you want to change your answer. (If your first choice is the car, he will reveal either goat at random 50% of the time.) Should you change your answer?

The End

Answers:

1) 13.4%

2) -5.26 bets

3) Yes!