BIG TODO that we need for putting entropy in the "weird.tex" paper. We need to know an interesting QM-generalization of "zero-phase" entropy. It is necessarily not Von Neumann since pure states are always zero Von Neumann entropy.

1 Classical Information Theoretic Entropy

We recall Shannon's classical entropy of the random variable P

$$H(P) := -\sum p_i \log(p_i)$$

and relative entropy, or Kullback–Leibler divergence, of a model random variable R with observations chosen from P

$$D_{KL}(P||R) := -\sum p_i \log \frac{r_i}{p_i}.$$

$H(P)$ is the average amount of "surprise" when using P, or the expected P information

$$H(P) = \mathbb{E}_P[-\log P].$$

For relative entropy we have another expected value appearing, the cross entropy, or expected R information

$$H(P||R) := \mathbb{E}_P[-\log R].$$

We see how these are related in

$$D_{KL}(P||R) = H(P||R) - H(P) \geq 0$$

where the last inequality is Gibbs' inequality. It is a result of Gibbs that equality holds exactly when $R = P$.

1.1 Standard Results

Lemma 1. *Let $f : X \to Y$ be a surjective map of finite random variables. Then $H(X) \geq H(Y)$. Furthermore, the following are equivalent*

1. *The fibers of f each have at most one nonzero probability.*
2. *The fibers of f each have zero entropy.*
3. *$H(X) = H(Y)$.*

Proof. When f is a bijection, $H(X) = H(Y)$, and we are done.

We build up non-bijective f as a series of maps

$$(A_0 = X) \to A_1 \to \cdots \to A_{k-1} \to (A_k = Y)$$

where $|A_i| = |A_{i-1}|$ - 1. In this way, we have a proof by induction on the size of Y. That is, it suffices to prove the case where $|Y| = |X|$ - 1 and apply it to the above $A_\bullet$ chain.

In this case there is a unique $y \in Y$ with non-singleton preimage. That y has exactly two unique preimages which we will call x_1 and x_2.

$$\{x_1, x_2\} \to \{y\}$$

Let the probabilities of seeing x_1 and x_2 be p_1 and p_2 respectively. The case where p_1 and p_2 are both zero results in $H(X) = H(Y)$ with the fiber condition satisfied, and we are done.

It remains to consider the case where at least one of the p_i is nonzero. The probability of seeing y is $p_1 + p_2$ which we will also call α for short. Let $q_i := p_i/\alpha$ and notice that $q_0 + q_1 = 1$. We finally can compare the contributions to the difference of H along the nontrivial fiber.

$$\begin{aligned}
H(X) - H(Y) &= -p_1 \log(p_1) - p_2 \log(p_2) + (p_1 + p_2) \log(p_1 + p_2) \\
&= -\alpha q_1 \left[\log(q_1) + \log(\alpha)\right] - \alpha q_2 \left[\log(q_2) + \log(\alpha)\right] + \alpha \log(\alpha) \\
&= \alpha(-q_1 \log(q_1) - q_2 \log(q_2)) \\
&\geq 0
\end{aligned}$$

which inductively shows that $H(X) \geq H(Y)$. The inequality is equality exactly when $\{q_1, q_2\} = \{0, 1\}$. Inductively this finishes the claim about the fibers. □

Lemma 2. *A nontrivial projection measurement decreases the entropy.*

Proof. Consider a finite set X, the disjoint union $X = X_1 \sqcup X_2$, and a measurement that projects onto X_2. This sends $p_i \to 0$ for each $i \in X_1$ and renormalizes the elements of X_2 by $\sigma = \sum_{i \in X_2} p_i$. Let $H(X)$ be the original entropy and let $H(\pi)$ be the entropy after the projection.

$$\begin{aligned}
H(X) - H(\pi) &= \sum_{i \in X} -p_i \log(p_i) + \sum_{i \in X_2} \frac{p_i}{\sigma} \log\left(\frac{p_i}{\sigma}\right) \\
&= \left(\sum_{i \in X_1} -p_i \log(p_i)\right) + \left(\sum_{i \in X_2} \left(-p_i + \frac{p_i}{\sigma}\right) \log(p_i)\right) - log(\sigma) \\
&\geq 0
\end{aligned}$$

Where the equality occurs exactly when $p_i = 0$ for all $i \in X_1$ and $\sigma = 1$, which is when the projection is trivial. □

Lemma 3. *We choose observations according to a distribution P and consider the Bayesian evidence towards another distribution R. The inference, on average, prefers P over R with equality exactly when $R = P$.*

Proof. Probabilities multiply so we need to work with logs to compute the expected values of the evidences. Then the Lemma becomes exactly Gibbs' inequality.

$$\mathbb{E}_P[-\log R] \geq \mathbb{E}_P[-\log P]$$

□