EECE 405/560

Information Theory and Perfect Secrecy

October 22, 2007

1 Information Theory

1.1 Definition of Entropy, and Mutual Information

For a single random variable X, which takes on values in the set $\mathcal{X}$:

$$H(X) = \sum_{x \in \mathcal{X}} \Pr[x] \log_2 \frac{1}{\Pr[x]}$$

For a pair of variables X, Y, the joint entropy:

$$H(X,Y) = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \Pr[x,y] \log_2 \frac{1}{\Pr[x,y]}$$

For a pair of variables X, Y, the conditional entropy:

$$H(X|Y) = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \Pr[x,y] \log_2 \frac{1}{\Pr[x|y]}$$

Note carefully!!! Both $H(X,Y)$ and $H(X|Y)$ are summed over $\Pr[x,y]$.

Given variables X and Y, the mutual information between them is

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

$$I(X;Y) = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \Pr[x,y] \log_2 \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$$

1.2 Identities

$$
\begin{aligned}
H(X) &\geq 0 && (1)\\
H(X) &\leq \log_2 |\mathcal{X}| \quad \text{equality when } X \text{ is uniform over } \mathcal{X} && (2)\\
H(X,Y) &= H(X) + H(Y|X) && (3)\\
H(X,Y) &= H(X) + H(Y) \quad \ldots \text{ if X,Y independent.} && (4)\\
H(X|Y) &= H(X) \quad \ldots \text{ if X,Y independent.} && (5)\\
I(X,Y) &\geq 0 \quad \ldots \text{ equality when X,Y independent.} && (6)\\
H(X|Y) &\leq H(X) && (7)\\
H(g(X)|X) &= 0 && (8)
\end{aligned}
$$

1.3 Proving the identies

Here are how you prove the above identities:

1. $H(X) \geq 0$ because $\log 1/p$ can never be negative.
2. Uses Jensen's inequality (below).
3. Just write out the definitions of $H(X)$ and $H(Y|X)$ and combine them to get the definition of $H(X,Y)$.
4. Follows from the definition of $H(X,Y)$ and the definition of independence: $\Pr[x,y] = \Pr[x]\Pr[y]$.
5. Follows from the previous two identities!
6. Uses Jensen's inequality.
7. Follows from previous identity: $I(X;Y) = H(X) - H(X|Y) \geq 0$.
8. Write out the definition of $H(g(X)|X)$.

Only two identities are complicated, using Jensen's inequality. For sake of completeness, here they are. But first, what is Jensen's inequality?

Definition 1.1 *A function $f()$ is* convex *if for values x_1, x_2 and any parameter $\alpha \in [0,1]$, $f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$.*

What does this mean? If we graph a convex function f(), and draw a chord between any two points on the graph, the chord will never fall below the curve. The chord is the set of points $\alpha f(x_1) + (1-\alpha) f(x_2)$ for all $0 \leq \alpha \leq 1$. The curve over that same interval is the set of all points $f(\alpha x_1 + (1-\alpha) x_2)$.

Theorem 1.1 Jensen's Inequality: *Suppose $f()$ is a convex function, and X is a random variable that takes on real numbers. Then $Ef(X) \geq f(EX)$.*

Proof: for a convex function $f()$ at any point x_0, we can always draw a line passing through $< x_0, f(x_0) >$ which never rises above the curve. Symbolically, there is a slope m such that

$$f(x) \geq m(x - x_0) + f(x_0)$$

Proving that there is such a slope is an exercise. But once we have that fact, we can simply take expectations of both sides:

$$\begin{aligned} Ef(X) &\geq Em(X - x_0) + Ef(x_0) \\ &\geq m(EX - x_0) + f(x_0) \\ &\geq m(EX - x_0) + f(x_0) \end{aligned}$$

Now, the trick is this: we can let x_0 be anything. So, set $x_0 = EX$, which is a specific real number. then $Ef(X) \geq m(EX - EX) + f(EX)$, and we're done.

Now that we have this inequality, we can use it to prove useful identities, since $f(p) = \log 1/p = -\log p$ is convex:

Identity 2:

Let $f(x) = 1/\Pr[x]$. Construct a random variable $\hat{X} = f(X)$. For any outcome x, let $\hat{x} = f(x)$. Then because the log function is *concave*:

$$\begin{aligned} H(X) &= \sum_{x\in\mathcal{X}} \Pr[x] \log \frac{1}{\Pr[x]} \\ &= \sum_{x\in\mathcal{X}} \Pr[x] \log \hat{x} \\ &= E[\log(\hat{X})] \\ &\leq \underbrace{\log E[\hat{X}]}_{\text{Jensen!}} \\ &= \log \left[\sum_{x\in\mathcal{X}} \Pr[x] \frac{1}{\Pr[x]}\right] = \log \left[\sum_{x\in\mathcal{X}} 1\right] \\ &= \log |\mathcal{X}| \end{aligned}$$

Identity 6:

Let $f(x, y) = \Pr[x] \Pr[y] / \Pr[x, y]$. Construct the random variable $Z = f(X, Y)$. Since the log is concave, $-\log x$ is convex:

$$\begin{aligned} I(X;Y) &= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr[x, y] \log_2 \frac{\Pr[x, y]}{\Pr[x] \Pr[y]} \\ &= -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr[x, y] \log_2 \frac{\Pr[x] \Pr[y]}{\Pr[x, y]} \qquad \text{Note the minus sign!} \\ &= \sum_{z\in\mathcal{Z}} \Pr[z] \cdot -\log z \\ &= E[-\log(Z)] \\ &\geq \underbrace{-\log E[Z]}_{\text{Jensen!}} \\ &= -\log \left[\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr[x, y] \frac{\Pr[x] \Pr[y]}{\Pr[x, y]}\right] = -\log \left[\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr[x] \Pr[y]\right] \\ &= -\log 1 = 0. \end{aligned}$$

2 Perfect secrecy

2.1 Notation

We have a set of plaintexts $\mathcal{P}$, a set of keys $\mathcal{K}$, and a set of ciphertexts $\mathcal{C}$. A plaintext and key are chosen at random, so P and K are random variables. $C = \texttt{Encrypt}(P, K)$ is also a random variable.

We will denote a plaintext by x (because p is overused), and a ciphertext by y. Our encryption and decryption algorithms are $y = \texttt{Encrypt}(x, k)$ and $x = \texttt{Decrypt}(y, k)$. For a fixed key, $\texttt{Decrypt}(\texttt{Encrypt}(x, k), k) = x$.

It is possible that more than one key maps x to y. For convenience, we define the set of keys D_{yx} like so: $D_{yx} = \{k \in \mathcal{K} : y = \texttt{Encrypt}(x, k)\}$. This is the set of all keys that map a given x to a given y. This notation allows us to succinctly express things like $\sum_{k \in Dyx} \Pr[k]$, the probability that a key is chosen that maps x to y. This happens to be $\Pr[y|x]$.

Because the key is chosen independently from the plaintext, $\Pr[k, p] = \Pr[k] \Pr[p]$. Together, $\Pr[k]$ and $\Pr[p]$ and the cipher algorithm give us a distribution on C:

$$\begin{aligned}
\Pr[y] &= \sum_x \Pr[x] \cdot \Pr[y|x] \\
\Pr[y|x] &= \sum_{k \in Dyx} \Pr[k] \\
\Pr[y] &= \sum_x \sum_{k \in Dxy} \Pr[x] \Pr[k] \\
&= \sum_{(x,k):y=\texttt{Encrypt}(x,k)} \Pr[x] \Pr[k]
\end{aligned}$$

2.2 Definition of perfect secrecy

The cipher has *perfect secrecy* if $\Pr[x] = \Pr[x|y]$. Intuitively, that means that an intercepted ciphertext provides no extra information about what x could be—our estimates of the probabilities are unchanged.

There are three equivalent definitions of perfect secrecy:

$$\begin{aligned}
\Pr[x] &= \Pr[x|y] \\
\Pr[x, y] &= \Pr[x] \Pr[y] \\
\Pr[y] &= \Pr[y|x]
\end{aligned}$$

2.3 Perfect secrecy and information

For any cipher, regardless of perfect secrecy, we have the following:

$$\begin{aligned}
H(P,K,C) &= H(P,K) + \underbrace{H(C|P,K)}_{0} = H(P,K) \\
&= H(P) + H(K) \qquad \ldots \text{ because P, K are independent} \\
H(P,K,C) &= H(K,C) + \underbrace{H(P|K,C)}_{0} = H(K,C) \\
H(P,K,C) &= H(P,C) + H(K|P,C)
\end{aligned}$$

We have $H(C|P,K) = 0$ because given a plaintext and key, there is no remaining uncertainty about the ciphertext. Likewise for $H(P|C,K)$.

Perfect secrecy gives us one extra rule: $H(P,C) = H(P) + H(C)$. This follows because $\Pr[x,y] = \Pr[x]\Pr[y]$ — the plaintext and ciphertext are independent. So combining the above rules we have

$$\begin{aligned}
H(P,K,C) &= H(P) + H(C) + H(K|P,C) \\
H(P,K,C) &= H(K,C) = H(C) + H(K|C) \\
\therefore H(P) + H(C) &= H(K|C) - H(K|P,C) + H(C) \\
H(P) &= H(K|C) - H(K|P,C) \\
&\leq H(K|C) \\
&\leq H(K)
\end{aligned}$$

This is the big result of perfect secrecy: the plaintext entropy can not be larger than the key entropy. Or in English: a cipher can only have perfect secrecy if on average, the key is at least as long as the message!

2.4 A cipher with perfect secrecy

Consider a shift cipher where all keys are chosen with equal probability 1/26. This is a lousy cipher in general, as you can easily guess even small messages: $\texttt{Encrypt}(x,k) = \texttt{`LIPPS ASVPH'}$. You can simply try all 26 keys.

However, if the message is only one letter, the eavesdropper is stuck: the ciphertext letter does not help us guess the plaintext letter. In fact, if the plaintext is one letter long, this cipher has perfect secrecy. For longer messages, you need to choose a new random shift for each letter—confirming our result that the key needs to be as long as the message.

This is a form of the *one-time pad*, a cipher in which random data is taken from a pad, used to encrypt each symbol, then never used again. Usually we see this described in binary: your plaintext is a bit string $b_1 b_2 b_3 \ldots b_n$, and we generate a random key $k_1 k_2 k_3 \ldots k_n$ by flipping n fair coins. The ciphertext is $c_i = b_i \oplus k_i$, where "$\oplus$" denotes the exclusive-or operation.

Rarely do people ever use one-time pads, because they require such a large key.