Information Theory and Security

Prakash Panangaden
McGill University

First Canada-France Workshop on Foundations and Practice of Security
Montréal 2008

Why do we care?

- Security is about controlling the flow of "information."
- We need to consider "adversaries" and analyze what they can do.
- Adversaries may have access to lots of data and perform statistical analysis.
- Thus we need to think probabilistically.

What is information?

- A measure of uncertainty.
- Can we really analyze it quantitatively?
- What do the numerical values mean?
- Is it tied to “knowledge”?
- Is it subjective?

Uncertainty and Probability

Suppose that you have a distribution

$$p_1 = \frac{1}{n}, \ldots, p_n = \frac{1}{n}$$

This is clearly very uncertain.

The other end

Consider a probability distribution like:

$$p_1 = 1, p_2 = 0, \ldots, p_n = 0.$$

We have a lot more “information.”

Conveying Information

Suppose that we want to convey the results of an election. There are 5 politicians running: Barak, Hillary, John, Mike and Maggie. It would normally take 3 bits to convey the result.

Suppose that the probabilities of winning are:

$$B : \frac{1}{2}, H : \frac{1}{4}, J : \frac{1}{8}, Mi, Ma : \frac{1}{16}$$

We can encode the results as:

$$B : 0, H : 01, J : 001, M_i : 0001, M_a : 0000$$

Which uses only $1\frac{7}{8}$ bits on the average.

What do we want?

We want a definition that satisfies the following conditions:

For a point distribution the uncertainty is 0

For a uniform distribution the uncertainty is maximized.

When we combine systems the uncertainty is **additive**

As we vary the probabilities the uncertainty changes

continuously

Entropy

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log_2 p_i$$

- $H(0, 0, \ldots, 1, 0, \ldots, 0) = 0$
- $H(\frac{1}{n}, \ldots, \frac{1}{n}) = \log_2 n.$
- Clearly continuous.

Are there other candidates?

Entropy is the unique continuous function that is:

- maximized by the uniform distribution
- minimized by the point distribution
- additive when you combine systems
- and ....

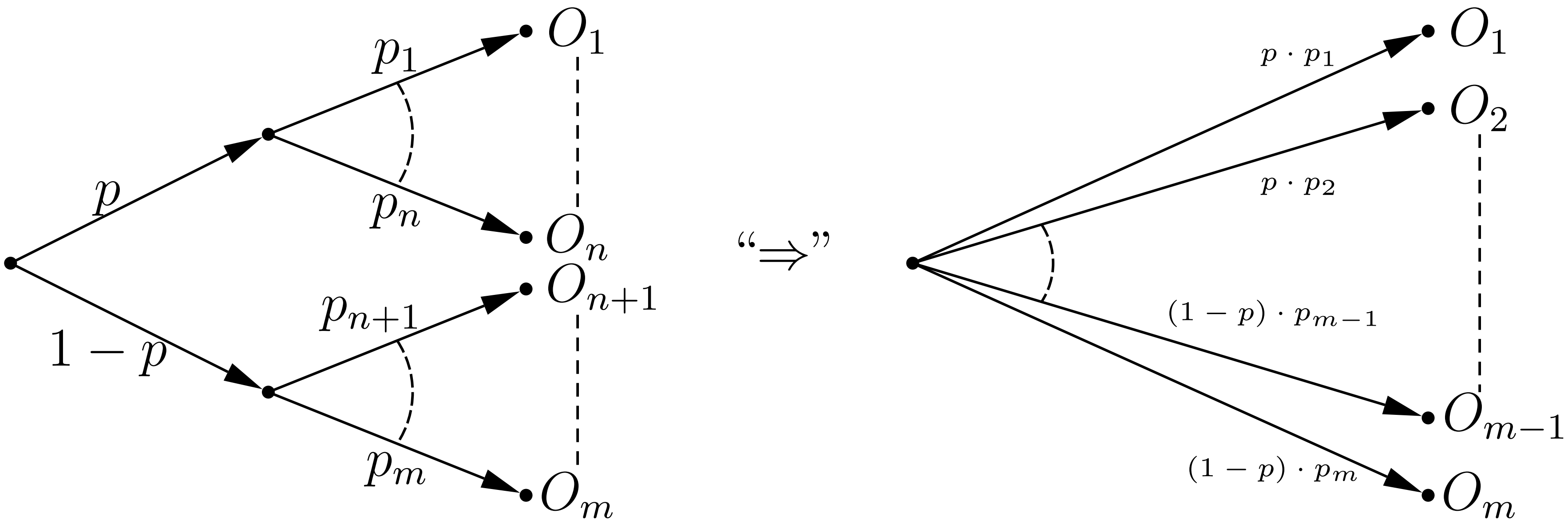
Grouping

$$H_m(p_1, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) +$$

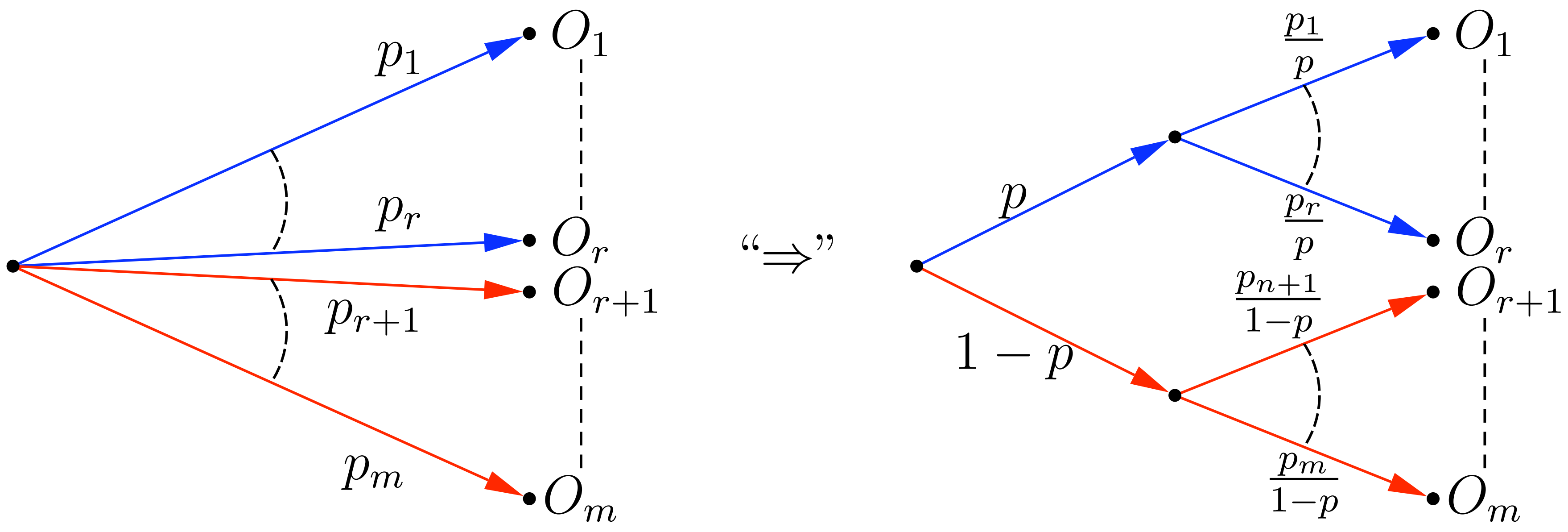
$$(p_1 + p_2)H_2(\frac{p_1}{p_1 + p + 2}, \frac{p_2}{p_1 + p_2})$$

What does this mean?

A picture of grouping 1



Grouping Picture 2



What does it tell us?

If you have a distribution $p(s)$ on a set S,
you can define a code such that it takes $H(p)$
bits on the average to encode the members of the set.

What we really care about

- In security we want to know how to extract secret information from readily available data.
- We want to measure how information (uncertainty) about one quantity conveys information about another.

Random variables

A discrete probability space is a finite set Ω equipped with a probability distribution

$$Pr : \Omega \to [0,1]$$

satisfying

$$\sum_{\omega \in \Omega} Pr(\omega) = 1.$$

A *random variable* X is a function from Ω to a finite set S.

A random variable induces a distribution on S via:

$$Pr_X(s \in S) = Pr(\{\omega : X(\omega) = s\})$$

$$Pr(X = s) = Pr(\{\omega : X(\omega) = s\})$$

Entropy of a Random Variable

$$H(X) = -\sum_{s \in S} Pr(X = s) \log_2 Pr(X = s)$$

We will just write $p(s)$ for $Pr(X = s)$ if the context is clear.

Joint Entropy

Consider a pair of random variables X, Y
taking values in sets $\mathcal{X}, \mathcal{Y}$
with a joint distribution $p(x, y)$.

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y)$$

Nothing new, yet!

Conditional Entropy

$H(Y|X = x)$ is the entropy of the random variable Y *given* that you know that X is x.

The conditional entropy is just the weighted sum:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

$$H(Y|X) \leq H(Y)$$

The Chain Rule

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

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

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

But

$$H(X|Y) \neq H(Y|X)$$

Mutual Information

The reduction in the uncertainty of one RV given another.

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

Recall, from the last slide, this means:

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

Hence, $I(X;Y) = I(Y;X)$

How far apart are distributions ?

We want a “distance” between distributions.

$$KL(p \mapsto q) = n \sum_{s \in S} p(s)[\log_2 p(s) - \log_2 q(s).$$

Recall that it takes $H(p)$ bits to describe a set distributed according to p. What if we used q instead?

It would require $H(p) + KL(p \mapsto q)$ bits.

Relative Entropy

The Kullback-Leibler distance is often called *relative entropy.*

Suppose $S = \{a, b\}$ and $p(a) = \frac{1}{2} = p(b)$ while $q(a) = \frac{1}{4}$ and $q(b) = \frac{3}{4}$.

$$KL(p \mapsto q) = 0.2075 \text{ and } KL(q \mapsto p) = 0.1887.$$

Relative Entropy and Mutual Information

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log\{\frac{p(x,y)}{p(x)p(y)}\}$$

which is equal to:

$$KL(p(x,y) \mapsto p(x)q(x))$$

A measure of how far you are from independence!

Some basic properties

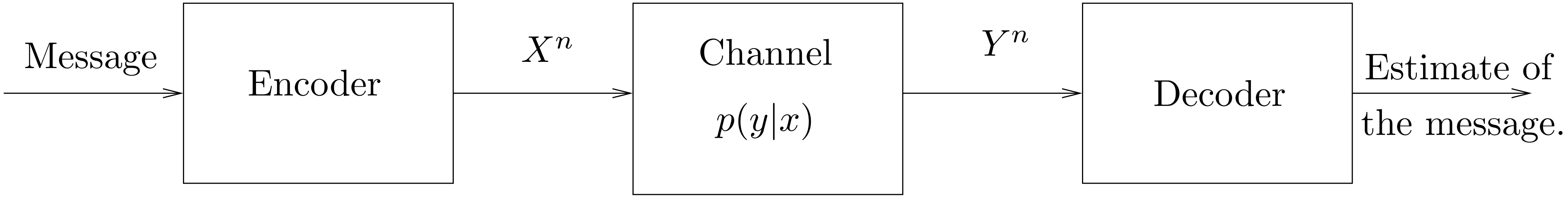
There are chain rules for mutual information and relative entropy.

$$KL(p \mapsto q) \geq 0.$$

hence

$$I(X;Y) \geq 0.$$

Channels



A typical channel.

How well can we estimate the intended message if the channel is noisy?

Channel Capacity

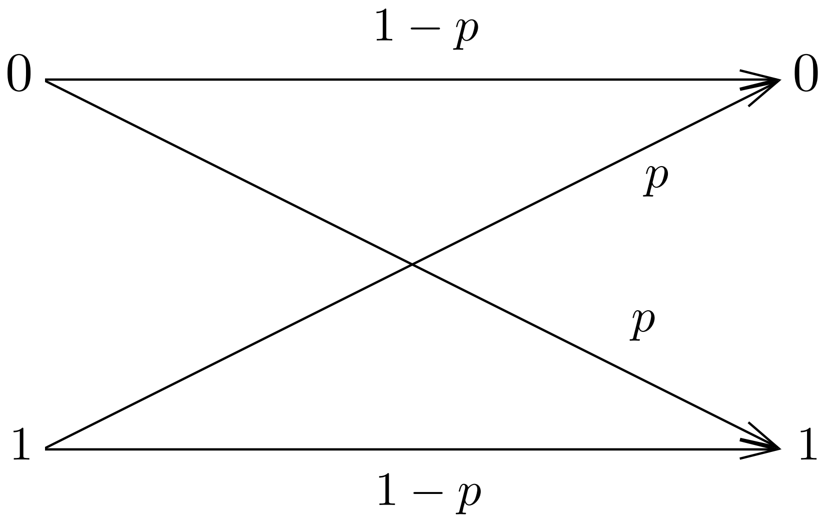
We want some way of measuring how well we can estimate the message based on what we receive.

How about $I(X;Y)$?

But this depends on the input distribution!

$$C = \max_{p(x)} I(X;Y).$$

Binary Symmetric Channel



With probability p the bit is flipped.

$$C = 1 - H(p)$$

Coding Theorem

Informal version!

All rates below the capacity are achievable.

There is some coding so that one can send bits with no loss of information as long as the transmission rate is below the capacity.

Capacity and Security

- We want to view protocols as channels: they transmit information.
- We would like our channels to be as bad as possible in transmitting information!
- Catuscia Palamidessi's talk: Channel capacity as a measure of anonymity.

Capacity of What?

- Ira Moskowitz et. al. studied the capacity of a covert channel to measure how much information could be leaked out of a system by an agent with access to a covert channel.
- We are viewing the protocol itself as an abstract channel and thus adopting channel capacity as a quantitative measure of anonymity.

Basic Definitions

- $\mathcal{A}$ is the set of anonymous actions and A is a random variable over it; a typical action is “a”.
- $\mathcal{O}$ is the set of observable actions and O is a random variable over it; a typical action is “o”
- $p(a,o) = p(a) * p(o|a)$
- $p(o|a)$ is a characteristic of the protocol; we design this
- $p(a)$ is what an attacker wants to know.

Anonymity Protocol

- An anonymity protocol is a channel $(\mathcal{A}, \mathcal{O}, p(.|.))$
- The loss of anonymity is just the channel capacity

Sanity Check

- To what does capacity 0 correspond?
- It corresponds precisely to strong anonymity, i.e. to the statement that A and O are independent.

Relative Anonymity

- Sometimes we **want** something to be revealed; we do not want this to be seen as a flaw in the protocol.
- We need to conditionalize everything.

Conditional Mutual Information

- Suppose that we want to reveal R
- For example, in an election we want to reveal the total tallies while keeping secret who voted for whom.
- Since we want to reveal R by design we can view it as an additional observable.

The *conditional mutual information* $I(A;O|R)$ is
$I(A;O|R) = H(A|R) - H(A|R,O)$.

Let $(\mathcal{A}.\mathcal{O}, p(\cdot|\cdot))$ be an anonymity protocol.
Let R be a random variable defined by a set of values $\mathcal{R}$
and a probability distribution $p(r|a,o)$.

The relative loss of anonymity (conditional channel capacity)
of the protocol with respect to R is
$C|R = \max_{p(a)} I(A;O|R)$.

An Example: Elections

- The actions are of the form “i votes for c”
- Suppose that there are two candidates, “c” and “d”; clearly we want to reveal the number of votes for “c” [everyone votes for exactly one candidate].
- Then the values of R are exactly the number of votes for “c.”
- The observable event is a scrambled list

Partitions

- This is a very special case: the hidden event, i.e. the complete description of who voted for whom, determines the value “r” of R.

- The observable event also completely determines “r”.

- Thus each value “r” produces partitions of the hidden values and of the

Let $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be an anonymity protocol and R a random variable with $\mathcal{R} = \{r_1, \ldots, r_n\}$. If R partitions both A and O, then the protocol matrix breaks into block diagonal form.

Each block is itself a channel with capacity C_i.
The conditional capacity
of the overall channel is bounded by
the maximum value of the C_is.

Computing the Capacity

- There is no simple (analytic) formula for the channel capacity.
- There are various special symmetric cases known; see the paper.
- Recently Keye Martin has shown that channels can be ordered as a domain and that capacity is Scott continuous on this domain. These results have been extended by Chatzikokolakis with Keye.

Conclusions

- Information theory is a rich and powerful way to analyze probabilistic protocols.
- A wealth of work to be done given how hard it is to compute anything exactly.
- All kinds of beautiful mathematics: convexity theory, domain theory in addition to traditional information theory.