# ILLC Project Course in Statistical Learning Theory

Mathias Winther Madsen
mathias.winther@gmail.com

Institute for Logic, Language, and Computation
University of Amsterdam

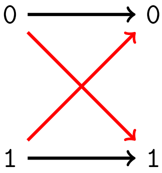
January 2015

# Information Theory

**Problem**

We are provided with a communication channel for binary information which tends to transmit as much as 3 in every 1000 bits wrongly (1 as 0, or 0 as 1).

If we feed a sequence of 1000 bits into this channel, how many different sequences might come out?




# Information Theory

$$\Phi(k,t) = \binom{t}{0} + \binom{t}{1} + \binom{t}{2} + \cdots + \binom{t}{k}$$

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 1 | | | | | | |
| | | | | | 1 | | 1 | | | | | |
| | | | | 1 | | 2 | | 1 | | | | |
| | | | 1 | | 3 | | 3 | | 1 | | | |
| | | 1 | | 4 | | 6 | | 4 | | 1 | | |
| | 1 | | 5 | | 10 | | 10 | | 5 | | 1 | |
| 1 | | 6 | | 15 | | 20 | | 15 | | 6 | | 1 |

$$\Phi(2,4) = 11$$

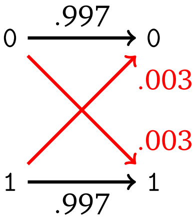
| | | |
|---|---|---|
| `...x` | `..xx` | `x.x.` |
| `..x.` | `.x.x` | `xx..` |
| `.x..` | `x..x` | `....` |
| `x...` | `.xx.` | |

# Information Theory

## Problem

We are provided with a communication channel which transmits each individual bit wrongly with probability 0.003.

If we feed a sequence of 1000 bits into this channel, how many “reasonably likely” output sequences are there?




# Information Theory

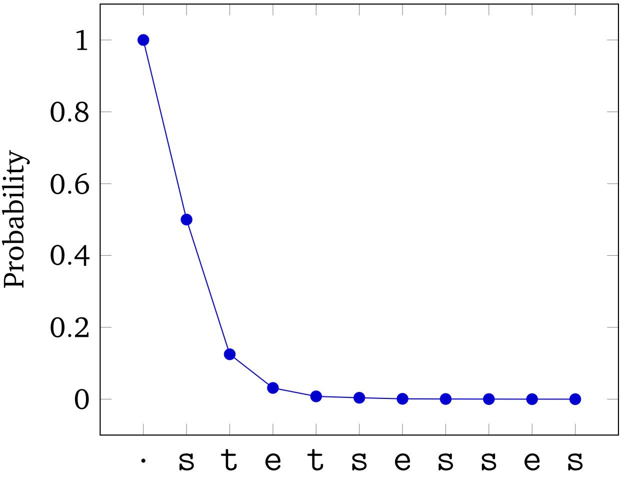
## Problem

With the point probabilities

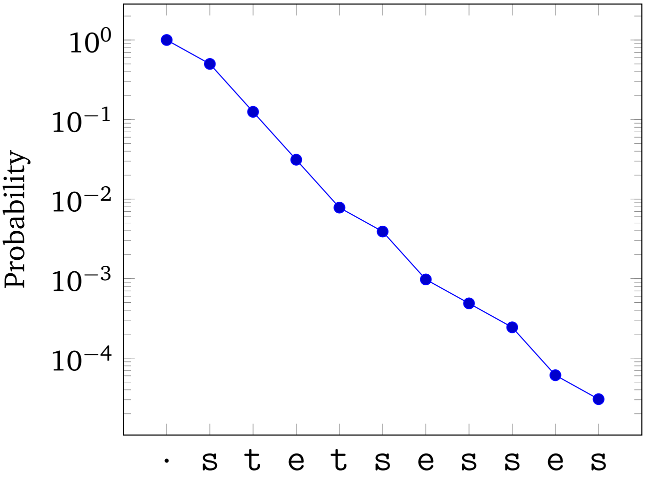
| $x$ | t | s | e |
|---|---|---|---|
| $p(x)$ | .25 | .50 | .25 |

1. What is Pr(`stetsesses`)?
2. What's the most probable sequence?

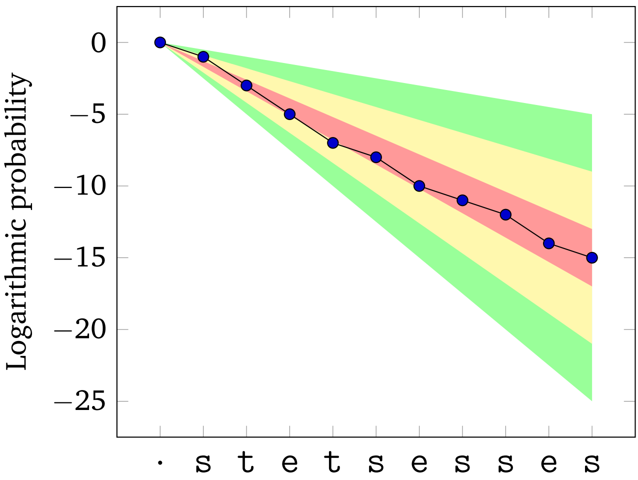
# Information Theory



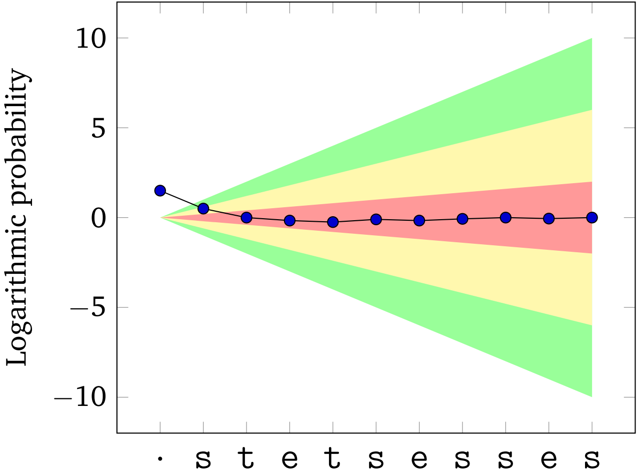
# Information Theory



# Information Theory



# Information Theory



# Information Theory

## Definition

The **entropy** of a random variable $X$ is

$$H = E\left[\log \frac{1}{p(X)}\right] = -E\left[\log p(X)\right].$$

## Definition

An $\varepsilon$**-typical sequence** of length $t$ is a sequence for which

$$\left|\log \frac{1}{p(x_1, x_2, \ldots, x_t)} - Ht\right| < \varepsilon.$$

# Information Theory

## The Weak Law of Large Numbers

For every $\varepsilon > 0$ and $\alpha > 0$ there is a $t$ such that

$$\Pr\left\{\left|\frac{\sum_{i=1}^{t} X_i}{t} - E[X]\right| > \varepsilon\right\} \le \alpha.$$

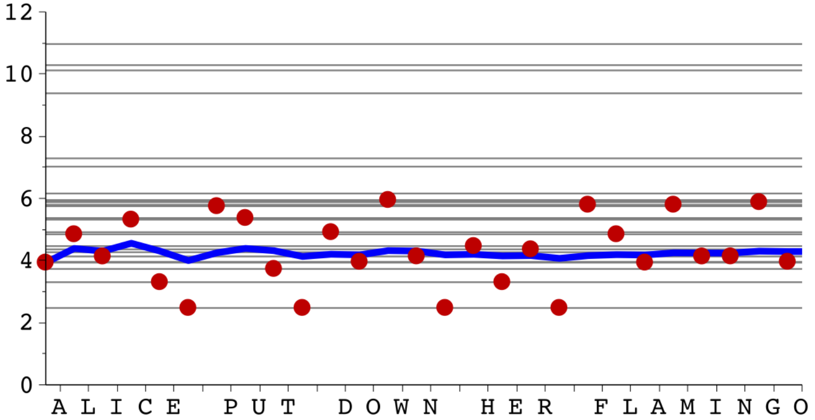
## The Asymptotic Equipartition Property

Eventually, everything has the same probability.

## The Source Coding Theorem

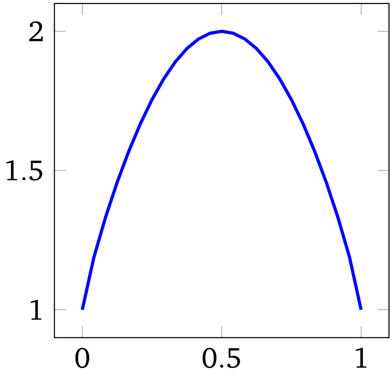
For large $t$, there are only $2^{Ht}$ sequences worth caring about.

# Information Theory



# Information Theory

$$\exp(H_2(q)) \;=\; \left(\frac{1}{q}\right)^{q} \left(\frac{1}{1-q}\right)^{1-q}.$$



# Information Theory

$$\exp(D(q\,||\,p)) \;=\; \left(\frac{p}{q}\right)^{q}\left(\frac{1-p}{1-q}\right)^{1-q} \;\leq\; \exp\left(-2\,(q-p)^2\right).$$