Wireless Connectivity:
An Intuitive and Fundamental Guide

Chapter 8: Information-Theoretic View on Wireless Channel Capacity

Petar Popovski

Connectivity Section

Department of Electronic Systems

petarp@es.aau.dk

Contributions to the slides:

Israel Leyva-Mayorga

Radoslaw Kotaba

Abolfazl Amiri

Alexandru-Sabin Bana

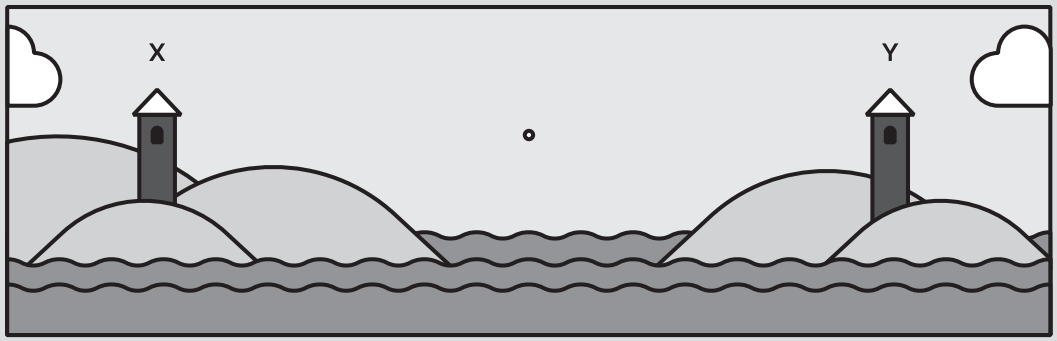
Robin J. WIlliams

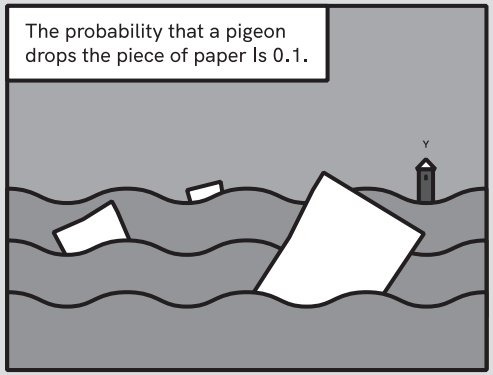
AALBORG UNIVERSITY

DENMARK

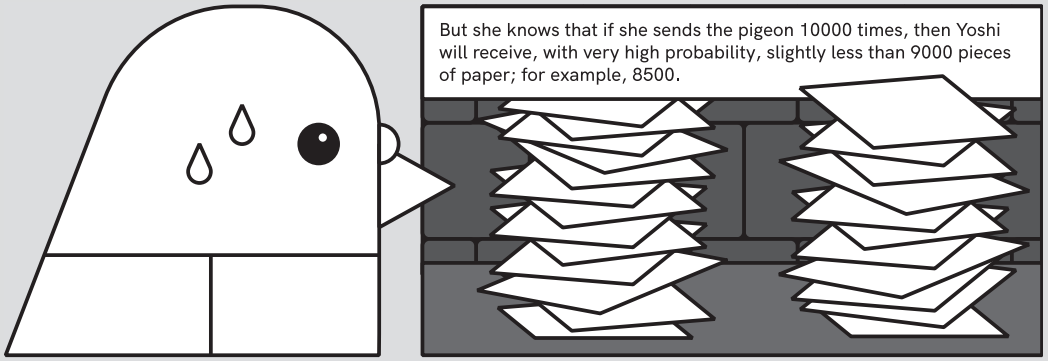
Modules

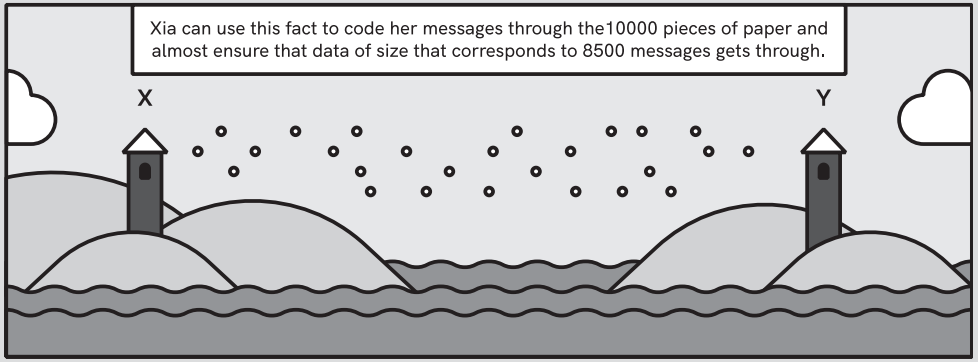
1. An easy introduction to the shared wireless medium
2. Random Access: How to Talk in Crowded Dark Room
3. Access Beyond the Collision Model
4. The Networking Cake: Layering and Slicing
5. Packets Under the Looking Glass: Symbols and Noise
6. A Mathematical View on a Communication Channel
7. Coding for Reliable Communication
8. **Information-Theoretic View on Wireless Channel Capacity**
9. Time and frequency in wireless communications
10. Space in wireless communications
11. Using Two, More, or a Massive Number of Antennas
12. Wireless Beyond a Link: Connections and Networks


X
Y

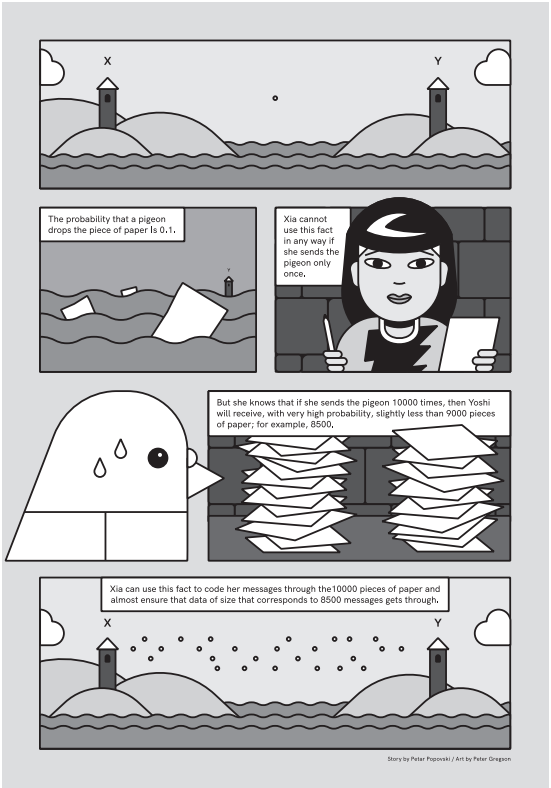

The probability that a pigeon drops the piece of paper Is 0.1.


Xia cannot use this fact in any way if she sends the pigeon only once.


But she knows that if she sends the pigeon 10000 times, then Yoshi will receive, with very high probability, slightly less than 9000 pieces of paper; for example, 8500.


Xia can use this fact to code her messages through the10000 pieces of paper and almost ensure that data of size that corresponds to 8500 messages gets through.
X
Y

Channel capacity



- We cannot do anything about reliability if the channel is used only once
- If the channel is used many times, then we can start to see statistical regularity
- Channel capacity is defined for asymptotically many uses of the channel

What will be learned in this chapter

- Why is the law of large numbers important to define channel capacity
- Typical sequences, perfectly reliable communication and channel capacity
- Mutual information
- The popular Shannon formula for a Gaussian channel
- Fading channels and capacity

Perfect reliability with nontrivial data rate

True reliability may trick us to add redundancy indefinitely (bringing the rate to 0)

Fortunately, this is not the case

- **Fundamental result of information-theory:** while, indeed, packet length needs to go to infinity, the amount of data increases at the same rate, while reliability approaches to a perfect one (zero probability of error)!
- The number of codewords M that can be selected grows as

$$M = 2^{lC}$$

with channel uses l and data rate per channel use $C = \frac{\log_2 M}{l} > 0$

The highest possible value for C is the **channel capacity**

The role of the law of large numbers

Simple illustration: Random bit generator producing 1 with p and 0 with $(1-p)$

Create a sequence of l i.i.d. random numbers

When the sequence is long, we **expect** to see lp 1's and $l(1-p)$ $0's$

- This is called a **typical sequence**
- Occurs with probability $P_T = p^{lp}(1-p)^{l(1-p)}$
- In terms of its entropy $P_T = 2^{\log_2\left(p^{lp}(1-p)^{l(1-p)}\right)} = 2^{-lH(p)}$
- As $l \to \infty$ we will almost surely observe one of them so

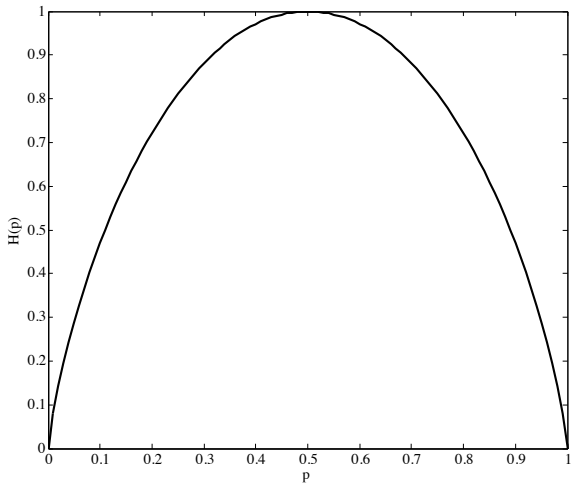
$$L_T \times P_T \approx 1 \Rightarrow L_T \approx \frac{1}{P_T} = 2^{lH(p)}$$

The law of large numbers (LLN) turns a **local property** into a **global property**

A digression into source coding

Bit vs. binary value

How many bits of information are required, on average, to represent the binary sequence produced by a random generator for which the probability of getting 1 is p?



The probability for a sequence to be typical is <1:

$$L_T P_T = 1 - P_e \qquad L_T \leq 2^{l(H(p)+\varepsilon)}$$

Encoding a typical sequence requires: $D_T = \lceil \log_2 L_T \rceil$ bits

Send typical compressed, send non-typical uncoded and add one bit to each to discriminate

$$\overline{D} = (1 - P_e)(1 + D_T) + P_e(l + 1) = 1 + (1 - P_e)(l(H(p) + \varepsilon) + 1) + P_e l$$

$$R = \frac{\overline{D}}{l} \approx H(p)$$

Perfectly reliable communication

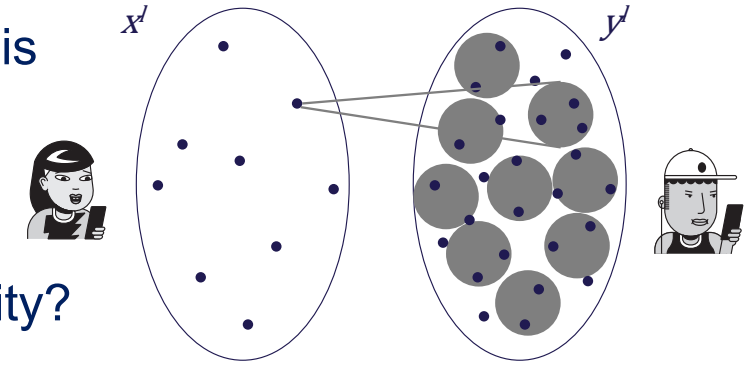
Consider a BSC with per-symbol error probability $p > 0$

Received signal: $y^l = x^l \oplus v^l$ where v^l is one of $L_T \approx 2^{lH(p)}$ typical noise vectors

For a given **input** x^l, there will be a cloud of L_T typical **outputs** y^l (gray area)

- At most $M = \frac{2^l}{2^{lH(p)}} = 2^{l(1-H(p))}$ non-overlapping codewords
- Can represent $\log_2 M$ bits, so the **maximal rate** is

$$R = \frac{\log_2 M}{l} = 1 - H(p)$$



Can we select M codewords such that
even errors (typical ones) do not introduce ambiguity?

- YES! How? Well…

The quest for the perfectly reliable code

Shannon's result does not provide a recipe for creating an optimal codebook:
It proves that it exists using a probabilistic argument

- Generate codewords one-by-one through coin flipping
- Then, gather them into codebooks $\{x_1^l, \ldots x_M^l\}$
- The error of the codebook is characterized by the overlap between codewords
  - This error tends to 0 as l goes to infinity

Such a codebook is just one of all possible codebooks, but…

By generating it this way it is an **average codebook**

- Then, there must exist at least one that is "better"

The maximal rate for which an arbitrarily low error probability can be guaranteed is known as **channel capacity...** **For BSC:** $C = 1 - H(p)$

Mutual information and its interpretations

Statistical properties of the single channel use reflect the channel capacity
This is the **local property** of the underlying channel

- Define a set of K possible inputs $x \in \{a_1, \dots a_K\}$ and a set of L possible outputs $y \in \{b_1, \dots, b_L\}$
- For a given $x = a_i$ the output is a random variable Y characterized by the conditional distribution $P(y|a_1)$ and entropy

$$H(Y|x = a_1) = \sum_{j=1}^{L} P(b_j|a_1) \log_2 \frac{1}{P(b_j|a_1)}$$

- The input is a random variable itself (having distribution $P(x)$)
- Then what is relevant is the **average** uncertainty (entropy) of Y given X

$$H(Y|X) = \sum_{i=1}^{K} P(a_i) H(Y|X = a_i)$$

Mutual information and its interpretations

Furthermore, given $P(y|x)$ and $P(x)$ we can determine the marginal

$$P(Y = b_j) = \sum_{i=1}^{K} P(a_i)P(Y = b_j|X = a_i)$$

which is necessary to calculate $H(Y)$

The difference between this marginal and conditional entropy is known as **mutual information**

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

Mutual information and its interpretations

In words:

We are given a channel $P(Y|X)$ that cannot change

We can decide on $P(X)$ (the only thing we **control directly**)

- Clearly, $P(X)$ and $P(Y|\mathrm{X})$ determine $P(Y)$
- $P(Y)$ is the only thing that we can **observe directly** so it is important that it provides as much information as possible about X (which we are really after)

Mutual information:

Information that a RV provides about another one

Hence, the goal is to find $P(X)$ that maximizes $I(X;Y)$

$$\sup_{\mathrm{P}(X)} I(X;Y) = C$$

For BSC, mutual information is maximized when $P(x) = \frac{1}{2}$

Mutual information in some practical communication setups

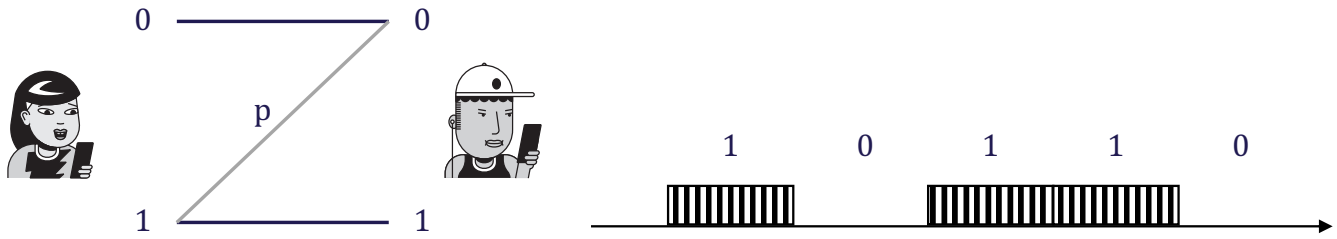
The codebook created by a fair coin toss is not always the right solution:

Consider a Z-channel where Xia transmits 0 by staying silent (high reliability) and 1 by transmitting 1's

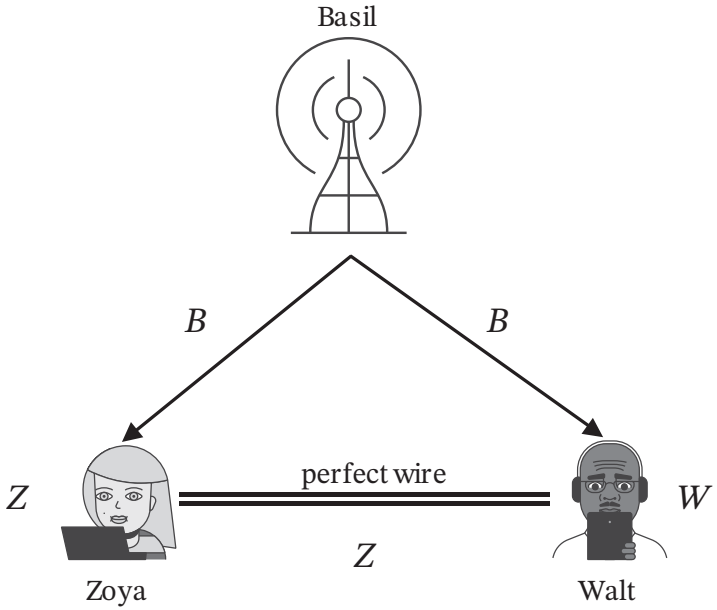
- Clearly, Xia should transmit 0 more often
- The input distribution that achieves the capacity of the Z-channel is

$$P(X=0) = 1 - \frac{1}{(1-p)\left(1-2^{\frac{H(p)}{1-p}}\right)}$$

The codewords should exhibit similar statistics in terms of their content (0's and 1's)

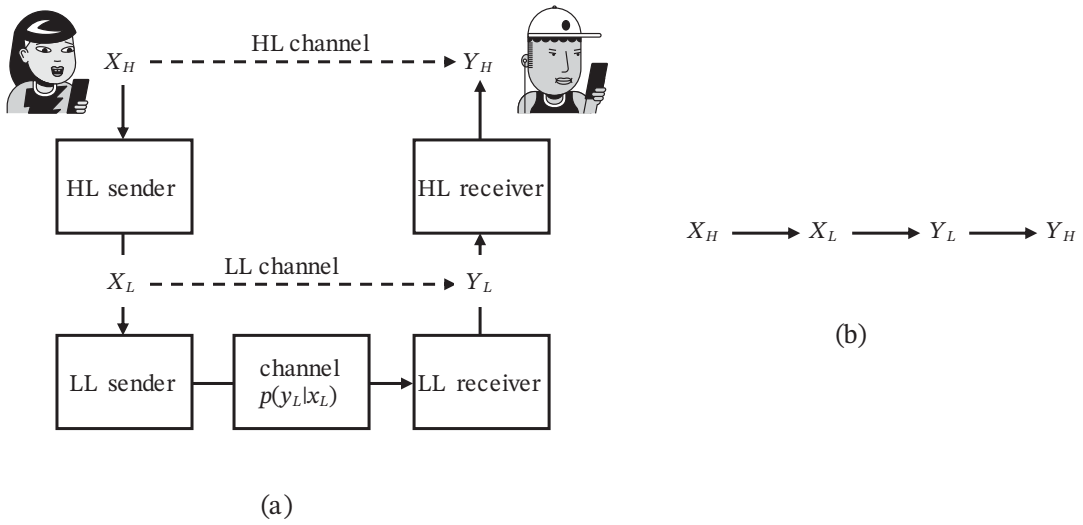


Mutual information in some practical communication setups



Broadcast channel with cooperation:

$$I(B;ZW) = I(B;W) + I(B;Z|W)$$



Layering and data processing inequality:

$$I(X_H;Y_H) \leq I(X_L;Y_L)$$

The Gaussian channel and differential entropy

The classical definition of entropy is not directly applicable when the random variables for communication are drawn from continuous distribution

For continuous RVs characterized by pdf $f(x)$ we talk about **differential entropy**

$$h(X) = -\int f(x) \log_2 f(x)\, \mathrm{d}x$$

Knowledge of $f(x)$ and $f(y|x)$ allows to determine $f(y)$ and $f(x|y)$

This, in turn, is used to obtain a conditional differential entropy $h(X|Y)$

Mutual information in continuous RVs

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

For continuous RVs it is possible to have $h(X) < 0$, but less intuitive than in the discrete case.

Worst case for noise: Entropy is the **highest** for Gaussian distribution among the RVs with fixed variance $E[|X|^2] = \sigma^2$

$$h(X_G) = \frac{1}{2}\log_2 2\pi e\sigma^2$$

The capacity of the Gaussian channel and the "Shannon formula"

What is the optimal input distribution?

The **signal power** is typically limited to P and is related to the variance of the signal

- When $E[|X|^2]$ is constrained, the highest entropy is observed for Gaussian RV, hence, ideally $X \sim \mathcal{N}(0, P)$
- When the noise is Gaussian with power $N = \sigma^2$:

  The observed Y is another Gaussian RV with statistics $\mathcal{N}(0, P + N)$

The maximal mutual information and, hence, **channel capacity** is

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right) = \frac{1}{2}\log_2(1 + \gamma)$$

To achieve twice that capacity, real and complex channels can be superimposed:

Recall the discussion about QPSK/BPSK

Achieving the Gaussian channel capacity

How to create codewords that will achieve the capacity?

Generate random codewords but this time following Gaussian distribution $\mathcal{N}(0, P)$ (instead of fair 50/50 coin)

As the length of a codeword $l \to \infty$, LLN guarantees that the average power

$$\frac{1}{l}\sum_{j=1}^{l}\left|x_{i,j}\right|^2 = P \text{ (observe the notation } x_i^l = \left[x_{i,1}x_{i,2} \ldots x_{i,l}\right])$$

Codeword x_i^l, with its associated **typical** noise cloud, has an l-dimensional volume

$$V_n \approx 2^{lh(n)} = (2\pi el\sigma^2)^{\frac{1}{2}} = (2\pi elN)^{\frac{1}{2}}$$

Since $Y \sim \mathcal{N}(0, P+N)$ due to uncertainty, the total volume of Y is

$$V_y \approx 2^{lh(y)} = \left(2\pi el(P+N)\right)^{1/2}$$

The maximum number of non-overlapping codewords is $M = \frac{V_y}{V_n}$ and

$$\frac{\log_2 M}{l} = \frac{\frac{l}{2}\log_2\frac{P+N}{N}}{l} = \frac{1}{2}\log_2(1+\gamma)$$

Interpretations of the "Shannon formula"

The capacity formula $C = \frac{1}{2}\log_2(1+\gamma)$ is often **misused**

Example: Gaussian channel $y = hx + n$ with SNR $\gamma = \frac{|h|^2 P}{N}$

To achieve capacity we need to send many symbols e.g. $l = 10000$

If h is constant, Xia can learn it and adapt the transmission to achieve the capacity

Else if h changes e.g. every 50 channel uses, then we could be tempted to write that the number of information bits is $50\sum_{i=1}^{200} C(\gamma_i)$

But the formulas are valid only due to LLN, which is **not fulfilled** here!

One could consider average SNR instead of individual γ's…

But it doesn't account for resources spent on learning h

If practical constellations are used (e.g. 16-QAM)
then they should be adapted per each symbol

Capacity of fading channels

We will focus on channels of the type $y = hx + n$
and consider different behaviors of h

We distinguish two main types of **fading**

1. **Slow fading** occurs when h, while random, stays constant for infinite or practically infinite number of channel uses
   If h is known, transmission can occur *at capacity*
2. ***Fast fading*** occurs when h changes, well, *fast*. **Example:** every L channel uses
   - **Most extreme case:** $L = 1$ and each symbol experience different fading
   - The time L during which channel stays constant is called **coherence time**

Regardless of L being finite or not, we can consider a communication over **infinitely** many periods V such that there are in total VL channel uses

Rule: we assume receiver knows h perfectly

Capacity with available CSI

We mean channel state information (CSI) available at the transmitter

Slow fading

Simplest case where Xia can transmit at a rate $C = \log_2(1 + \frac{|h|^2 P}{N})$

If transmissions occur over V blocks and Xia uses **constant** power, the average rate is

$$R_P = \frac{1}{V}\sum_{v=1}^{V} C\left(\frac{|h_v|^2 P}{N}\right) = \frac{1}{V}\sum_{v=1}^{V} C(\gamma_v)$$

Since h_v is a random variable, we can talk about the **distribution** of the SNR

Rayleigh fading is commonly considered

Both real and imaginary parts of h are uncorrelated Gaussian RVs

For a given P, the distribution of the SNR is **exponential** with mean GP

$$p(\gamma) = \frac{1}{PG} e^{-\frac{\gamma}{PG}}$$

Capacity without available CSI

We start with slow fading

There are 2 states: **"good"** and **"bad"** occurring with p_G and $p_B = 1 - p_G$

- During **"good"** states Xia can (successfully) transmit with rate

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

- During **"bad"** states no information is conveyed:

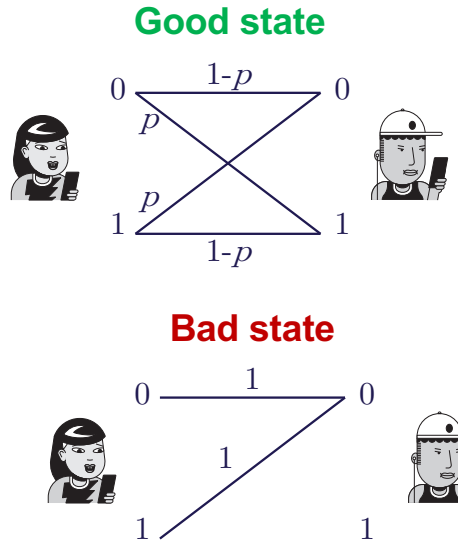
So, capacity is 0, right?

well…,the average throughput is still positive:

$$T = p_G\big(1 - H(p)\big)$$

Outage probability is $(1 - p_G)$

To combat this, Yoshi can employ **feedback**



Capacity without available CSI

Fast fading with $L = 1$: The situation changes significantly
The scenario is characterized by probabilities p_{ij}

If neither Xia nor Yoshi have CSI:

- Xia generates the codewords for the binary channel defined by these probabilities
  - With $p_G = 0.9$ and $p = 0.01$, the capacity is 0.7104 bits per channel use
  - **Observe:** 0 is more reliable it is expected to be used more often

Else if Yoshi knows CSI, he can distinguish between different "0s"

- The **capacity increases** $C = (1 - H(p\,))p_G = 0.8273$
- This is equal to the slow fading case **with feedback**

In ***fast fading***, Xia can achieve positive rate due the **law of large numbers**
The channel visits all states during the l channel uses of the codeword

Channel estimation and knowledge

Costly, but oftentimes **beneficial**

Accurate h is required for the fading channel for **Maximum Ratio Combining** (MRC)

Knowledge of is $|h|$ needed in AWGN for rate selection

Communication phases of a protocol:

1. **Estimation** with v pilot symbols $\hat{h}_v = h + w_v$
   More pilots = smaller variance of noise estimation
2. **Reporting** with r symbols
   SNR is not exactly reported – but *quantized* in K bits
   It indicates one of $R_1, \ldots, R_K$ possible coding rates
3. **Data communication** with rate adaptation

Channel estimation and knowledge

How large should v be?

$$y_v = \hat{h}_v^* y = (h^* + w_v^*)(hx + z) = |h|^2 x + n_{vx},$$

Note: n_{vx} is not exactly Gaussian, but can be assumed to be…
If one is **conservative** when selecting the rate

Effective data rate:

$$G_k = \frac{l}{v + r + l} R_k$$

Average long term goodput:

$$\bar{G} = \frac{l}{v + r + l} \sum_{k=1}^{K} P_k R_k \,,$$

where P_k is the probability that the k-th state occurs

Outlook and takeaways

- Law of Large Numbers is essential
  - It leads to **typical behavior**

- **"Optimal"** might be hard to find, but random can be pretty good

- Significance of mutual information

- Communication over slow and fast fading is fundamentally different

… and as usual, who knows what really matters