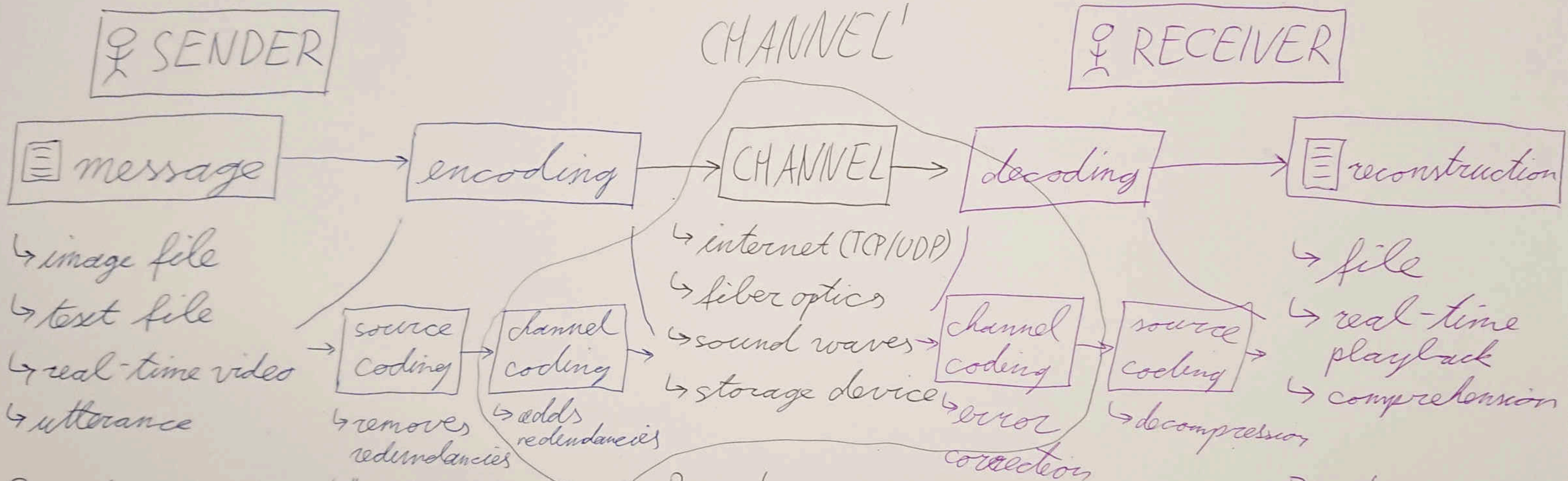
DATA COMPRESSION WITH DEEP PROBABILISTIC MODELS

Robert Bamler, University of Tuebingen

20 April 2021

Problem Setting: Communication over a channel



Properties:
- digital or analog
- contains redundancies

Properties
- digital or analog
- noisy or noise free
- finite transfer rate
  "channel capacity"

Properties
- lossy vs. lossless
- streaming ("progressive")
  or "bulk"
- seekable?

Goal: transmit message from S to R: fast + reliable

LOSSLESS COMRESSION I: SYMBOL CODES

Problem Setting

- communicate over a noise free channel
- sender has message $\underline{x}$, wants to transmit it losslessly to receiver in as few bits as possible
- encoder:

$$\underline{x} \longmapsto C^*(\underline{x}) \in \underbrace{\{0,1\}^*}_{\text{set of all bit strings of arbitrary length}}$$

Kleene star

- more generally: $C^*(\underline{x}) \in \{0, \ldots, B-1\}^*$ with $B \in \{2, 3, 4, \ldots\}$ ("B-ary code") (commonly: $B=2$)

Symbol Codes

- message $\underline{x}$ is a sequence of symbols x_i from a discrete alphabet $\mathcal{X}$:

$$\underline{x} = (x_1, x_2, \ldots, x_k) \equiv (x_i)_{i=1}^{k}$$

where $x_i \in \mathcal{X}\ \forall i$ and $k \in \mathbb{N}$ and $\mathcal{X}$ is finite (or countably infinite)

- encoder: $C^*(\underline{x}) = C(x_1) \| C(x_2) \| \ldots \| C(x_k)$

concatenation

↳ C is called the "code book"

↳ $C(x)$ is called the "code word for symbol $x \in \mathcal{X}$"

↳ Def: $\ell(x) :=$ length of $C(x)$ (i.e., number of bits)

Examples of Symbol Codes

1) Morse code: $B=3$ (dot, dash, pause)
2) UTF-8: $(B=2)$
   - ↳ $\mathcal{X} = \{$all UNICODE code points$\}$
   - ↳ $C(x)$ = UTF-8 representation of x
   - ↳ $\ell(x) \in \{8, 16, 24, 32\}$ (bits)

3) "Simplified game of Monopoly":
   - throw a pair of dice several times, after each time, write down their sum as a new symbol x;
   - for simplicity, let's use 3-sided dice

$$\Rightarrow \mathcal{X} = \{\underset{\substack{1+1}}{2},\ \underset{\substack{1+2,\\ 2+1}}{3},\ 4,\ 5,\ \underset{3+3}{6}\}$$

$$\boxed{C^*((2,6)) = \underbrace{101}\underbrace{10} = C^*((5,2))}$$

- possible code books:
  - ↳ $C^{(1)}(x)$ = binary representation of x
  - ↳ $C^{(2)}(x)$ = " $(x-2)$
  - ↳ $C^{(3)}(x)$ = " $(x-2)$, padded to consistent length
  - ↳ $C^{(4)}(x)$, $C^{(5)}(x)$: see table

x	$C^{(1)}(x)$	$C^{(2)}(x)$	$C^{(3)}(x)$	$C^{(4)}(x)$	$C^{(5)}(x)$
2	~~10~~	~~0~~	000	010	010
3	~~11~~	~~1~~	001	10	01
4	~~100~~	~~10~~	010	00	00
5	~~101~~	~~11~~	011	11	11
6	~~110~~	~~100~~	100	011	110

<u>Reminder:</u> We ultimately want to encode & decode a <u>sequence</u> of symbols, not just a single one (in as few bits as possible).