

# Definition: Sources and Codes

We start by investigating symbol codes: codes that encode a source  $P_X$ , one symbol at a time. Later on, we will also see codes that group the source symbols together into blocks.

## Definition: Symbol code

Let  $P_X$  be the distribution of a random variable  $X$  (with image  $\mathcal{X}$ ), and let  $\mathcal{A}$  be a finite set. A symbol code for the source  $P_X$  and with alphabet  $\mathcal{A}$  is an **injective function**  $C : \mathcal{X} \rightarrow \mathcal{A}^*$ .

Here,  $\mathcal{A}^* = \bigcup_{n \in \mathbb{N}} \mathcal{A}^n \cup \perp$ , and  $\perp$  is the empty string. That is,  $\mathcal{A}^*$  is the set of finite sequences of elements from  $\mathcal{A}$ : this operation on sets is called the **Kleene star**.

We often refer to the set of codewords,  $\mathcal{C} = \text{im}(C)$ , as code and leave the actual encoding function  $C$  implicit.

In many instances, the alphabet  $\mathcal{A}$  is fixed to be the set  $\{0, 1\}$  of size 2. In that case, we speak of a **binary symbol code**. The codewords of a binary code are simply binary strings.

## Definition: Codeword length

Let  $C : \mathcal{X} \rightarrow \mathcal{A}^*$  be an encoding function. For any  $x \in \mathcal{X}$ , the length  $\ell(C(x))$  of the codeword  $C(x)$  is the length of the sequence of symbols from  $\mathcal{A}$ . That is, if  $C(x) \in \mathcal{A}^k$ , then  $\ell(C(x)) = k$ .

For practical applications, it is important that the codewords are (on average) short: that way, the transmission or storage of a message is as efficient as possible.