

## Lecture 2

Instructor: Madhu Sudan

Scribe: Zachary Ziegler

## 1 Administrative Notes

- Sign up on Piazza if haven't already
- Sign up for scribing. If needed, double up after spring break
- PS1 due Fri 2/8
- Follow <http://people.seas.harvard.edu/~madhusudan/courses/Spring2019/>
- Start thinking about potential final projects

## 2 Formal Definition of Entropy

Let  $X$  be a random variable with probability distribution  $P_X$ . Last class we defined entropy informally as “the number of bits needed, in expectation, to convey  $X$ ”. Technically, this definition is incorrect, as demonstrated by the following example:

**Example 1.** Let  $X_1, \dots, X_{100} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p = 0.01)$ . According to the axioms introduced in Lecture 1,  $H(X_1, \dots, X_{100}) = \sum_{i=1}^{100} H(X_i)$  because each  $X_i$  is independent. One bit is needed to convey each  $X_i$ , so the RHS has value 100. However,  $p = 0.01$  is small, indicating that we could compress the joint set and convey the information in many fewer bits. This implies that under the previous definition of entropy  $H(X_1, \dots, X_{100}) < \sum_{i=1}^{100} H(X_i)$ , violating the axioms.

The correct definition, in words, is:

**Definition 2** (Entropy). Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_X$ . The *entropy* of  $X$  is the limit as  $n \rightarrow \infty$  of the number of bits needed, in expectation and on average, to convey the  $n$  iid samples of  $X$ .

To make this formal we introduce an encoder and decoder function. For  $X \in \Omega, \forall n$ ,

$$\begin{aligned} E_n : \Omega^n &\rightarrow \{0, 1\}^* \\ D_n : \{0, 1\}^* &\rightarrow \Omega^n \times \{?\} \\ \text{s.t. } \forall \omega \in \Omega^n & D_n(E_n(\omega)) = \omega \\ \forall \omega^{(1)} \neq \omega^{(2)} & E_n(\omega^{(1)}) \text{ not a prefix of } E_n(\omega^{(2)}) \end{aligned}$$

An encoder and decoder pair  $(E_n, D_n)$  satisfying these requirements is called a *valid pair*. Note that the prefix-free requirement is sufficient to ensure the mapping is invertible, but gives additional nice properties. Given these mappings, we define *entropy* formally as

$$H(x) \triangleq \lim_{n \rightarrow \infty} \left\{ \min_{(E_n, D_n) \text{ valid}} \left\{ \frac{1}{n} \mathbb{E}_{x \sim P_x^n} \left[ |E_n(x)| \right] \right\} \right\}$$

where  $|E_n(x)|$  denotes the length of the binary encoding.

### 3 Binomial Entropy Computation

While the previous discussion gives the operational definition, in practice we want to compute entropy directly from the distribution  $P_X$ . First, we consider the case  $X \sim \text{Bernoulli}(p)$ . According to the definition above, we need to consider  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$ . In this case, the sample  $(X_1, \dots, X_n)$  forms a binary sequence of length  $n$ . We use the following encoding procedure:

1. Alice sends Bob the number of ones in the sequence,  $k = \sum X_i$
2. Alice sends Bob the index of the correct binary sequence, among the  $\binom{n}{k}$  possibilities consisting of  $k$  ones (they have previous agreed on an ordering).

The number of bits to convey an integer  $a$  is  $\log a$ , therefore

$$\mathbb{E}_{x \sim P_x^n} \left[ |E_n(x)| \right] = \log n + \log \binom{n}{k}$$

By the weak law of large numbers,

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P \left( \left| \sum X_i - np \right| > \varepsilon \right) = 0$$

As the definition of entropy involves  $\lim_{n \rightarrow \infty}$ , it suffices in the following discussion to consider  $k = \sum X_i = np$  with the understanding that additional terms exist which go to 0 as  $n \rightarrow \infty$ .

Introducing *Stirling's approximation*,

**Definition 3** (Stirling's approximation). For large  $n$ ,  $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$

we conclude

$$\begin{aligned} \log \binom{n}{pn} &\approx h(p)n + o(\log n) \\ h(p) &= p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \end{aligned}$$

**Exercise 4.** Show that  $\log \binom{n}{pn} \approx h(p)n + o(\log n)$ .

In the limit  $n \rightarrow \infty$  the  $\log n$  terms disappear, and dividing by  $n$  per the entropy definition gives

$$H(X) \leq h(p)$$

To make this an equality we need to show  $H(X) \geq h(p) \forall (E_n, D_n)$ . For any encoding, the receiver needs to distinguish between  $\binom{n}{k}$  possible strings. Strings of length  $|E_n(x)| \leq \log \binom{n}{pn}$  only occur with probability  $\leq 2^{-t}$ . Thus strings of length  $|E_n(x)| \geq \log \binom{n}{pn}$  occur with probability  $\geq 1 - 2^{-t}$ . Thus for any valid encoding,

$$\lim_{n \rightarrow \infty} \mathbb{E}_{x \sim P_X^n} |E_n(x)| \geq \log \binom{n}{pn} - t$$

Dividing by  $n$  and using the same logic as before, this reduces to  $H(x) \geq h(p)$ . Combining with the previous result, we find  $H(x) = h(p)$ .

## 4 Multinomial Entropy Computation

Let  $\Omega = \{1, \dots, l\}$  and  $P_X = (P_1, \dots, P_l)$ , where  $P_i = \Pr[X = i]$ . Again we take  $n$  iid samples  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_X$ . For large  $n$ , with high probability any string has  $p_1 n$  1's,  $p_2 n$  2's, ...,  $p_l n$   $l$ 's. Any string with these counts are equally likely, leading to an expected compressed length,

$$\begin{aligned} \mathbb{E}_{x \sim P_x^n} \left[ |E_n(x)| \right] &= l \log n + \log \binom{n}{p_1 n, p_2 n, \dots, p_l n} \\ &= h(p_1, \dots, p_l) n + o(\log n) \end{aligned}$$

With  $h(p_1, \dots, p_l) \triangleq \sum_{i=1}^l p_i \log \frac{1}{p_i}$ . Thus for a general (finitely supported discrete) distribution,

$$H(X) = \sum_{i \in \Omega} \Pr[x = i] \log \frac{1}{\Pr[x = i]}$$

**Exercise 5.** Similarly to the Bernoulli case, show that  $H(X) \geq h(p_1, \dots, p_l)$  in the multinomial case to formally conclude the proof.

**Exercise 6.** Suppose a fixed size encoding is used, but a fraction of error  $\gamma$  is allowed. Precisely, change the definition for a valid pair to require

$$\text{error} = \Pr[D_n(E_n(x)) \neq x] \leq \gamma$$

Show that the definition of  $H(X)$  does not change. Further, show that  $\gamma$  is exponentially small in  $n$ .

## 5 Asymptotic Equipartition Principle

In both the Bernoulli and multinomial cases we saw that the optimal encoding consisted of finding a subset of  $\Omega$  over which the distribution of encodings was uniform. The Asymptotic Equipartition Principle (AEP) generalizes this notion formally:

**Definition 7** (Asymptotic Equipartition Principle). For every finite set  $\Omega$ , every  $P_X$ , and every  $\varepsilon > 0$ , for sufficiently large  $n$ ,

$$\begin{aligned} \exists S \subseteq \Omega^n \text{ s.t. } & 1. \Pr_{x \sim P_X^n} [x \notin S] \leq \varepsilon \\ & 2. \forall \omega \in S \frac{1}{|S|^{1+\varepsilon}} \leq \Pr_{x \sim P_X^n} [x = \omega] \leq \frac{1}{|S|^{1-\varepsilon}} \end{aligned}$$

**Exercise 8.** Identify the correspondence between parts 1. and 2. of the Asymptotic Equipartition Principle with the Bernoulli and Multinomial entropy derivations from class.

**Exercise 9.** Prove  $|S| \approx 2^{H(X)n}$ .