01.12.2020. Lecture 13.

1. Conditional entropy.

We started the lecture with a reminder :

Definition 1. Let (α, β) be jointly distributed random variables, with

$$p_{ij} = \text{Prob}[\alpha = 1_i \ \& \ \beta = b_j].$$

For each value b_j we have a conditional distribution on the values of α with probabilities

$$p'_i = \text{Prob}[\alpha = \alpha_i \,|\, \beta = b_j] = \frac{\text{Prob}[\alpha = 1_i \ \& \ \beta = b_j]}{\text{Prob}[\beta = b_j]}.$$

This conditional distribution has Shannon's entropy ; we denote it $H(\alpha \,|\, \beta = b_j)$.

Definition 2. We define the entropy of α conditional on β as the following average value :

$$H(\alpha \,|\, \beta) := \sum_j \text{Prob}[\beta = b_j] \cdot H(\alpha \,|\, \beta = b_j).$$

We proved several properties of *conditional entropy* :

- $H(\alpha \,|\, \beta) \geq 0$
- $H(\alpha \,|\, \beta) = 0$ if and only if α is a deterministic function of β
- $H(\alpha, \beta) = H(\alpha \,|\, \beta) + H(\beta)$
- $H(\alpha \,|\, \beta) \leq H(\alpha)$
- $H(\alpha \,|\, \beta) = H(\alpha)$ if and only if α and β are independent

Interpretation of the conditional entropy. In the class we discussed an interpretation of the quantity $H(\alpha|\beta)$ as the optimal compression rate for the following encoding scheme with a *helper*. Let (α_i, β_i), $i = 1, \ldots, n$ be a sequence of independent and identically distributed pairs. We assume that both Sender and Receiver know the values of β_i, while only Sender knows α_i. In this setting, Sender should send to Receiver a message that allows to the latter to compute the values of α_i. The question is how long should be the message so that Receiver gets the values of $(\alpha_1, \ldots, \alpha_n)$ with a high (close to 1) probability. Note that a message of length $n \cdot H(\alpha) + o(n)$ bits is enough even without the "helper" β_i. It turns out that with the helping information the length of the message can be reduced to $H(\alpha_i|\beta_i) \cdot n + o(n)$ bits.

2. Mutual information.

Definition 3. We define the information in α on β as

$$I(\alpha : \beta) := H(\beta) - H(\alpha \,|\, \beta).$$

In the class we proved several properties of the *mutual information* :

— $I(\alpha : \beta) \geq 0$

— $I(\alpha : \beta) = 0$ if and only if α and β are independent

— $I(\alpha : \beta) = I(\beta : \alpha) = H(\alpha) + H(\beta) - H(\alpha, \beta)$

— $I(\alpha : \beta) \leq H(\alpha)$

— $I(\alpha : \beta) \leq H(\beta)$

Definition 4. We define the information in α on β conditional on γ as

$$I(\alpha : \beta \,|\, \gamma) := H(\beta \,|\, \gamma) - H(\alpha \,|\, \beta, \gamma).$$

3. The fundamental relations between different entropic quantities.

For a pair of jointly distributed random variables (α, β) we have the following entropic quantities : $H(\alpha)$, $H(\beta)$, $H(\alpha, \beta)$, $H(\alpha|\beta)$, $H(\beta|\alpha)$, $I(\alpha : \beta) = I(\beta : \alpha)$. These values are not totally independent. Indeed, given $H(\alpha)$, $H(\beta)$, $H(\alpha, \beta)$ we can compute the conditional entropies and the mutual information :

$$\begin{array}{rcl} H(\alpha|\beta) & = & H(\alpha, \beta) - H(\beta) \\ H(\beta|\alpha) & = & H(\alpha, \beta) - H(\alpha) \\ I(\alpha : \beta) & = & H(\alpha) + H(\beta) - H(\alpha, \beta). \end{array}$$

Similarly, given the quantities $H(\alpha|\beta)$, $H(\beta|\alpha)$, and $I(\alpha : \beta)$ we can compute the values of unconditional entropies :

$$\begin{array}{rcl} H(\alpha) & = & H(\alpha, \beta) + H(\beta|\alpha) \\ H(\beta) & = & H(\alpha, \beta) + H(\alpha|\beta) \\ H(\alpha, \beta) & = & I(\alpha : \beta) + H(\alpha|\beta) + H(\beta|\alpha). \end{array}$$

Notice that in both cases we need to know three parameters to determine all other entropic quantities for a (α, β).

For a triple of jointly distributed random variables (α, β, γ) we have much more different non-trivial entropic quantities :

$$\begin{array}{l} H(\alpha), H(\beta), H(\gamma), \\ H(\alpha, \beta), H(\alpha, \gamma), H(\beta, \gamma), H(\alpha, \beta, \gamma), \\ H(\alpha|\beta), H(\beta|\alpha), H(\alpha|\gamma), H(\gamma|\alpha), H(\beta|\gamma), H(\gamma|\beta), \\ H(\alpha|\beta, \gamma), H(\beta|\alpha, \gamma), H(\gamma|\alpha, \beta), \\ H(\alpha, \beta|\gamma), H(\beta|\alpha, \gamma), H(\gamma|\alpha, \beta), \\ I(\alpha : \beta), I(\alpha : \gamma), I(\beta : \gamma), \\ I(\alpha, \beta : \gamma), I(\alpha, \gamma : \beta), I(\alpha : \beta, \gamma), I(\alpha : \beta|\gamma), I(\alpha : \gamma|\beta), I(\beta : \gamma|\alpha). \end{array}$$

Sometimes it is useful to introduce one more quantity, "the mutual information of the triple,"

$$I(\alpha : \beta : \gamma) := H(\alpha) + H(\beta) + H(\gamma) - H(\alpha, \beta) - H(\alpha, \gamma) - H(\beta, \gamma) + H(\alpha, \beta, \gamma).$$

Again, these entropic quantites are not totally independent : it is enough to know seven "basic" quantities $H(\alpha), H(\beta), H(\gamma), H(\alpha,\beta), H(\alpha,\gamma), H(\beta,\gamma), H(\alpha,\beta,\gamma)$ to compute all other values of conditional entropies and mutual informations. In the class we discussed that, alternatively, all these entropic quantities can be computed given the other seven parameters,

$$H(\alpha|\beta,\gamma), H(\beta|\alpha,\gamma), H(\gamma|\alpha,\beta), I(\alpha,\beta:\gamma), I(\alpha,\gamma:\beta), I(\alpha:\beta,\gamma), I(\alpha:\beta:\gamma).$$

We used a Venn-like diagram to visualize the relations between all entropic quantities for α, β, γ.

08.12.2020. Lecture 14.

1. Discussion of the homework. We discussed the exercise of the homework :

— Let (α,β) be a pair of jointly distributed random variables. Then $I(\alpha:\beta) = H(\alpha)$ if and only if α is a deterministic function of β.

— Let (α,β,γ) be a triple of jointly distributed random variables. The conditional mutual information is defined as $I(\alpha:\beta|\gamma) = H(\beta|\gamma) - H(\beta|\alpha,\gamma)$. Then

$$\begin{array}{lrcl} (a) & I(\alpha:\beta|\gamma) & = & H(\alpha|\gamma) + H(\beta|\gamma) - H(\alpha,\beta|\gamma), \\ (b) & I(\alpha:\beta|\gamma) & = & H(\alpha,\gamma) + H(\beta,\gamma) - H(\alpha,\beta,\gamma) - H(\gamma), \\ (c) & I(\alpha:\beta|\gamma) & = & I(\beta:\alpha|\gamma), \\ (d) & I(\alpha:\beta|\gamma) & = & \sum_k \mathrm{Prob}[\gamma = c_k] \cdot I(\alpha:\beta|\gamma = c_k) \end{array}$$

From the last property it is easy to obtain that $I(\alpha:\beta|\gamma) \geq 0$ for all (α,β,γ).

— we constructed a distribution (α,β,γ) such that $I(\alpha:\beta:\gamma) < 0$. More specifically, we constructed a joint distribution (α,β,γ) such that

$$\begin{array}{l} H(\alpha) = H(\beta) = H(\gamma) = 1, \\ H(\alpha,\beta) = H(\beta,\gamma) = H(\alpha,\gamma) = 2, \\ H(\alpha,\beta,\gamma) = 2. \end{array}$$

For such a distribution we have $I(\alpha:\beta:\gamma) = -1$.

We also proved the following fact :

— If α and β are distributed on $\{a,b,c,d,e\}$ (a domain with five elements) and $\mathrm{Prob}[\alpha \neq \beta] < 1/2$, then $H(\alpha|\beta) < 2$.

2. Non-basic information inequality. We used the Venn-like diagram and the basic information inequalities to prove that for all jointly distributed (α,β,γ)

$$2H(\alpha,\beta,\gamma) \leq H(\alpha,\beta) + H(\alpha,\gamma) + H(\beta,\gamma).$$

We observed that this inequality can be reduced to the sum of three "basic" inequalities

$$\begin{array}{rcl} I(\alpha:\beta) & \geq & 0, \\ I(\alpha:\gamma|\beta) & \geq & 0, \\ I(\beta:\gamma|\alpha) & \geq & 0. \end{array}$$

2. Information theoretical cryptography. We started discussing the Vernam cipher and its optimality for a symmetric encoding/decoding scheme (to be completed in the next lecture).

15.12.2020. Lecture 15.

1. One-time pad encryption technique.

We proved that the *Vernam cipher* is secure : the mutual information between the secret message and the encoded message is equal to zero. Then we proved that this scheme is essentially optimal (i.e., the length of the secret key can be reduced without losing the security) :

Theorem [Shannon]. Let $(\mathbf{k}, \mathbf{m}, \mathbf{e})$ be a triple of jointly distributed random variables (in our context $\mathbf{k}$ is the secret key, $\mathbf{m}$ is the secret message, and $\mathbf{e}$). Assume that these random variables satisfy the following three conditions :

$$\begin{cases} H(\mathbf{e}|\mathbf{k},\mathbf{m}) = 0 & \text{(the cypher text can be obtained given the secret text and the secret key)} \\ H(\mathbf{m}|\mathbf{k},\mathbf{e}) = 0 & \text{(the original text can be reconstructed given the cypher text and the secret key)} \\ I(\mathbf{e}:\mathbf{m}) = 0 & \text{(the cypher text without the key gives no information on the original text).} \end{cases}$$

Then $H(\mathbf{k}) \geq \mathbf{m}$ (the size of the secret key must be at least the same as the entropy of the message).

In the class we proved this theorem. Moreover, we proved that the conclusion $H(\mathbf{k}) \geq \mathbf{m}$ remains true even if we omit the second condition.

2. Secret sharing.

We discussed in the class the notion of a *perfect secret sharing*, with simple classical examples. In this setting, a *secret* is a random variable S_0 (usually a uniform distribution on some finite set), which is understood as a distribution on possible values of a secret keys. We want to "distribute" this secret among k participants of the project so that (i) every "authorized" group of participants could reconstruct uniquely the value of S_0, and (ii) every "non-authorized" group of participants gets no information about the secret. Technically, this means that we include the random variable S_0 in a joint distribution $(S_0, S_1, \ldots, S_k)$ (where S_0 is the secret and $S_1 \ldots S_k$ are *shares* assigned to each participant) so that the conditions (i) and (ii) are satisfied.

Example 1. Let $k = 3$ and let us require that only all 3 participants know the secret S_0, and every group of less than 3 participants gets no information on S_0. In case when S_0 is a uniform distribution on $\{0,1\}^n$, there is a simple scheme satisfying the conditions (i) and (ii) : the "shares" of the secret S_1, S_2, S_3 are independent and uniform distribution on $\{0,1\}^n$, and S_0 is the bitwise XOR of them,

$$S_0 = S_1 \oplus S_2 \oplus S_3.$$

Then it is easy to verify that

$$(i)\ H(S_0 \,|\, S_1, S_2, S_n) = 0$$

and

$$(ii)\ H(S_0 \mid S_i, S_j) = H(S_0).$$

Example 2. Let $k = 3$ and let us require that every *two* participants know the secret but every *single* participant does not know the secret. In other words, we require that

$$H(S_0|S_1, S_2) = H(S_0|S_1, S_3) = H(S_0|S_2, S_3) = 0,$$
$$H(S_0|S_1) = H(S_0|S_2) = H(S_0|S_3) = H(S_0).$$

In this example we assume that S_0 is a uniform distribution on $\mathbb{Z}/p\mathbb{Z}$ for a prime number p.

In this setting the secret sharing can be implemented as follows : we fix some (non-zero) elements $x_1, x_2, x_3 \in \mathbb{Z}/p\mathbb{Z}$ and define the joint distribution $(S_0, S_1, \ldots, S_k)$ as follows : let a, b be independent uniformly chosen elements in $\mathbb{Z}/p\mathbb{Z}$, and respectively

$$Q(x) = a + bx$$

be a randomly chosen polynomial of degree less than 2 (again, over the field $\mathbb{Z}/p\mathbb{Z}$). We let $S_0 = a$ and

$$S_i := Q(x_i) \text{ for } i = 1, 2, 3.$$

Then it can be shown that S_0 is uniquely determined by any two shares S_i, S_j. The same time, S_0 has no mutual information with one single S_i.

Example 3. Now we generalize the previous example for the setting with $k > 3$ participants. Let us choose a parameter (threshold) t between 1 and k and require that (i) every group of at east t participants knows the secret, and (ii) every group of less than t participants gets no information about the secret. For simplicity, we assume that S_0 is a uniform distribution on $\mathbb{Z}/p\mathbb{Z}$ for a prime number p (in what follows we assume that $k < p$).

In this setting the secret sharing can be implemented in Shamir's scheme : we fix some elements $x_0, x_1, \ldots, x_k \in \mathbb{Z}/p\mathbb{Z}$ and define the joint distribution $(S_0, S_1, \ldots, S_k)$ as follows : let $a_0, \ldots a_{t-1}$ be independent uniformly chosen elements in $\mathbb{Z}/p\mathbb{Z}$, and respectively

$$Q(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_{t-1} x^{t-1}$$

be a randomly chosen polynomial of degree less than t (again, over the field $\mathbb{Z}/p\mathbb{Z}$), and

$$S_i := Q(x_i) \text{ for } i = 0, 1, \ldots, k.$$

Then it can be shown that

$$(i)\ H(S_0 \mid S_{i_1}, \ldots, S_{i_t}) = 0$$

for all $1 \leq i_1 < \ldots i_t \leq k$ (given t different points (x_i, S_i) on the graph of the polynomial $Q(x)$, we can reconstruct the coefficients of $Q(x)$ and therefore compute $S_0 = Q(x_0)$) and

$$(ii)\ H(S_0 \mid S_{i_1}, \ldots, S_{i_{t-1}}) = \log p = H(S_0)$$

(if we know only $t-1$ points $(x_i, Q(x_i))$ on the graph of the polynomial, than *all* values of $Q(x_0)$ are possible and equiprobable).

3. Kolmogorov complexity

Definition. Let $L : \{0,1\}^* \to \{0,1\}^*$ be a (possibly partial) computable function. We define $C_L(x) = \min\{|p| \ : \ L(p) = x\}$. (With the convention that minimum of the empty set is infinity.) The function L is often called a *description method* or a *decompressor*, and $C_L(x)$ is the complexity of x with respect to this decompressor.

Definition. Let L_1, L_2 be two computable functions. We say that L_1 is *better* than L_2 as a decompressor, if there exists a number Const such that for all strings $x \in \{0,1\}^*$

$$C_{L_1}(x) \le C_{L_2}(x) + \text{Const}.$$

In the class we proved the following statement :

Theorem. There exists a decompressor (a partial computable function) L_{opt} that is better than any other decompressor, i.e., for all computable L there exists a constant d such that for all binary strings x

$$C_{L_{opt}}(x) \le C_L(x) + d.$$

Such a decompressor L_{opt} is called *optimal*.

Remark. There exist infinitely many optimal decompressors. However, they are equivalent to each other in the following sense. If L_{opt} and L'_{opt} are tow optimal decompressors, that there exists a constant Const such that for all binary strings x

$$|C_{L_{opt}}(x) - C_{L'_{opt}}(x)| \le \text{Const}.$$

We fix some optimal decompressor L_{opt} and denote in what follows

$$C(x) := C_{L_{opt}}(x).$$

The value of $C(x)$ is called *Kolmogorov complexity* of x.

In a similar way we define *conditional* Kolmogorov complexity :

Definition. Let $L : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}^*$ be a (possibly partial) computable function. We define $C_L(x \,|\, y) = \min\{|p| \ : \ L(p,y) = x\}$. (With the usual convention that minimum of the empty set is infinity.) The function L is often called *decompressor* with a condition, and $C_L(x \,|\, y)$ is the complexity of x given y with respect to this decompressor.

Definition. Let L_1, L_2 be two computable functions of two arguments. We say that L_1 is *better* than L_2 as a decompressor, if there exists a number Const such that for all strings $x, y \in \{0,1\}^*$

$$C_{L_1}(x \,|\, y) \le C_{L_2}(x \,|\, y) + \text{Const}.$$

Theorem. There exists a decompressor (a partial computable function) L^c_{opt} that is better than any other decompressor, i.e., for all computable L and for all binary strings x

$$C_{L^c_{opt}}(x \,|\, y) \le C_L(x \,|\, y) + O(1).$$

Such a decompressor L^c_{opt} is called *optimal*.

We fix some optimal decompressor L^c_{opt} and denote in what follows

$$C(x \,|\, y) := C_{L^c_{opt}}(x \,|\, y).$$

The value of $C(x)$ is called *conditional Kolmogorov complexity* of x given y.

Basic properties of Kolmogorov complexity.

In the class we proved several properties of Kolmogorov complexity :

— there exists a constant d_1 such that for all binary strings x

$$C(x) \le |x| + d_1;$$

— there exists a constant d_2 such that for all binary strings x

$$C(xx) \le |x| + d_2;$$

— there exists a constant d_3 such that for all binary strings x

$$C(xx) \le C(x) + d_3;$$

— there exists a constant d_4 such that for all binary strings x

$$C(\underbrace{xx\ldots x}_{n}) \le |x| + 2\log n + d_4;$$

— for every n there exists a binary string x of length n such that $C(x) \ge n$;

Proposition. There exist constants d_1, d_2 such that for all binary strings x of length n with pn *zeros* and $(1-p)n$ *ones* we have

$$C(x) \le \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right) n + d_1 \log n + d_2.$$

Definition. Information in x on y is defined as $I(x : y) := C(y) - C(y \,|\, x)$.

We discussed (without a proof) the following theorem :

Theorem [Kolmogorov–Levin]. (a) There exist integer numbers c_1, c_2 such that for all binary strings x, y

$$\big| C(xy) - C(x) - C(y \,|\, x) \big| \le c_1 \log N + c_2,$$

where $N = C(x) + C(y)$ and xy denotes the concatenation of x and y.

(b) The mutual information is symmetric up to an additive logarithmic term :

$$|I(x : y) - I(y : x)| \le c_1 \log N + c_2,$$

where $N = C(x) + C(y)$.