Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 10
Scribed by Giorgio Magri — T 3/13

1 Asymptotic Equipartition Property

The following theorem is a consequence of the Weak Law of Large Numbers which will be crucial in what follows.

Theorem 1. ASYMPTOTIC EQUIPARTITION PROPERTY. *Consider a sequence $X_1, X_2,$ of i.i.d. random variables with finite range distributed accordingly to a probability mass function p; then:*

$$-\frac{1}{n}\log p(X_1, , X_n) \xrightarrow{p} H(p) \tag{1}$$

in words: the random variable $-\frac{1}{n}\log p(X_1, , X_n)$ converges in probability to the entropy $H(p)$.

Proof. Consider the new random variables $Y_1, Y_2,$ defined by $Y_i \doteq -\log p(X_i)$. Since the X_i are i.i.d., then the Y_i are i.i.d. too, given that functions of independent random variables are also independent random variables. The Weak Law of Large Numbers thus ensures that (2) holds.

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mathbf{E}_p\big[Y\big] \tag{2}$$

Note that:

$$\begin{aligned} -\frac{1}{n}\log p(X_1, , X_n) &= -\frac{1}{n}\log\prod_{i=1}^{n} p(X_i) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i && (3)\\ H(p) &= -\sum_{x\in Im(X)} p(x)\log p(x) = \mathbf{E}_p\big[Y\big] && (4) \end{aligned}$$

Thus, claim (1) thus immediately follows by replacing (3) and (4) in (2). □

2 Typical sets

Definition 1. Consider a probability distribution $p(x)$ over a finite set $\mathcal{X}$ and arbitrary $\epsilon > 0$ and $n \in \mathbb{N}$; the set $A_\epsilon^n(p)$ defined as follows:

$$A_\epsilon^n(p) \doteq \left\{\mathbf{x} = (x_1, , x_n) \in \mathcal{X}^n \;\middle|\; \frac{1}{2^{n[H(p)+\epsilon]}} \le p(x_1, , x_n) \le \frac{1}{2^{n[H(p)-\epsilon]}}\right\} \tag{5}$$

is called the *typical set* for p corresponding to ϵ and n. We will often write just A_ϵ^n instead of $A_\epsilon^n(p)$, when no confusion arises.

Theorem 2. *The following properties hold:*

1. $A_\epsilon^n(p) = \left\{\mathbf{x} \in \mathcal{X}^n \;\middle|\; \big| -\frac{1}{n}\log p(\mathbf{x}) - H(p)\big| > \epsilon\right\}$.
2. $\mathbf{P}_p\big(A_\epsilon^n(p)\big) > 1-\epsilon$, *for n large enough.*
3. $\big|A_\epsilon^n(p)\big| \le 2^{n[H(p)+\epsilon]}$ *for every n.*
4. $\big|A_\epsilon^n(p)\big| \ge (1-\epsilon)2^{n[H(p)-\epsilon]}$, *for n large enough.*

Proof. [1] The proof of the first claim of the theorem amounts to the following trivial chain of implications:

$$
\begin{aligned}
\mathbf{x} \in A_\epsilon^n(p) \quad &\iff \quad \frac{1}{2^{n[H(p)+\epsilon]}} \leq p(x_1, , x_n) \leq \frac{1}{2^{n[H(p)-\epsilon]}} \\
&\iff \quad -n\big[H(p)+\epsilon\big] \leq \log p(\mathbf{x}) \leq -n\big[H(p)-\epsilon\big] \\
&\iff \quad H(p)-\epsilon \leq -\frac{1}{n}\log p(\mathbf{x}) \leq H(p)+\epsilon \\
&\iff \quad \Big| -\frac{1}{n}\log p(\mathbf{x}) - H(p)\Big| > \epsilon
\end{aligned}
$$

[2] The following chain of inequalities holds for every $\epsilon, \delta \in (0,1)$ and every n sufficiently large:

$$
\begin{aligned}
\mathbf{P}_p\big(A_\epsilon^n(p)\big) \quad &= \quad \mathbf{P}_p\Big\{\mathbf{x} \in \mathcal{X}^n \;\Big|\; \big| -\frac{1}{n}\log p(\mathbf{x}) - H(p)\big| > \epsilon\Big\} \\
&\leq \quad 1-\delta
\end{aligned}
$$

where in the first step I have used the first claim of the theorem and in the second step I have used the AEP. By setting $\delta = \epsilon$, we obtain the second claim of the theorem. [3] The proof of the third claim of the theorem amounts to the following chain of inequalities:

$$
\begin{aligned}
1 \quad &= \quad \sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x}) \\
&\geq \quad \sum_{\mathbf{x}\in A_\epsilon^n(p)} p(\mathbf{x}) \\
&\geq \quad \sum_{\mathbf{x}\in A_\epsilon^n(p)} \frac{1}{2^{n[H(p)+\epsilon]}} \\
&= \quad \big|A_\epsilon^n(p)\big| \frac{1}{2^{n[H(p)+\epsilon]}}
\end{aligned}
$$

where in the third step I have used the first claim of the theorem, namely the fact that $p(\mathbf{x}) \geq \frac{1}{2^{n[H(p)+\epsilon]}}$ for every $\mathbf{x} \in A_\epsilon^n(p)$. [4] The proof of the fourth claim of the theorem amounts to the following chain of inequalities:

$$
\begin{aligned}
1-\epsilon \quad &\leq \quad \mathbf{P}_p\big(A_\epsilon^n(p)\big) \qquad \text{for } n \text{ large enough} \\
&\leq \quad \sum_{\mathbf{x}\in A_\epsilon^n(p)} \frac{1}{2^{n[H(p)-\epsilon]}} \\
&= \quad \big|A_\epsilon^n(p)\big| \frac{1}{2^{n[H(p)-\epsilon]}}
\end{aligned}
$$

where in the first step I have used the second claim of the theorem and in the second step I have used the first claim, namely the fact that $p(\mathbf{x}) \leq \frac{1}{2^{n[H(p)-\epsilon]}}$ for every $\mathbf{x} \in A_\epsilon^n(p)$. □

3 Codes

Let $\mathcal{V}$ be a finite set, whose elements are called *symbols*; a *word* on $\mathcal{V}$ is any finite concatenation of symbols of $\mathcal{V}$; the set of all words is denoted by $\mathcal{V}^*$; the number of symbols concatenated in a word $\omega \in \mathcal{V}^*$ is called the *length* of ω and denoted by $\ell(\omega)$; for any two words $\omega_1, \omega_2 \in \mathcal{V}^*$, we say that ω_2 is a *prefix* of ω_1 iff there exists $\omega_3 \in \mathcal{V}^*$ such that $\omega_1 = \omega_2\omega_3$, i.e. ω_1 is the concatenation of ω_2 followed by ω_3; we will usually assume $\mathcal{V} = \{0, 1\}$. With this little background, we can now state the following crucial definition.

Definition 2. Consider a discrete random variable X with finite range $\mathcal{X}$ and probability distribution p. A *code* for X by means of an alphabet $\mathcal{V}$ is a function C of the following form:

$$C : \mathcal{X} \to \mathcal{V}^* \tag{6}$$

For each $x \in \mathcal{X}$, the string $C(x)$ is called the *codeword* corresponding to x with respect to the code C and the length of the word $C(x)$ is denoted by $\ell_C(x)$ (or often just by $\ell(x)$, when no confusion arises). A code is called *non-singular* iff it is an injective function, namely the following holds for every $x, x' \in \mathcal{X}$: if xx', then $C(x)C(x')$. A code is called *binary* if $\mathcal{V} = \{0, 1\}$; we will usually consider binary codes.The quantity $\ell(C)$ defined as follows:

$$\ell(C) \stackrel{\triangle}{=} \mathbf{E}_p\big[\ell_C(X)\big] = \sum_{x \in \mathcal{X}} p(x)\ell_C(x) \tag{7}$$

is called the *expected length* of the code C.

4 Compression via typical sets

Theorem 3. *Consider a sequence of i.i.d. random variables $X_1, , X_n$, with common finite range $\mathcal{X}$. For any $\epsilon > 0$ and any $n \in \mathbb{N}$ large enough, there exists a non-singular binary code $C_\epsilon : \mathcal{X}^n \to \{0, 1\}^*$ such that its expected length is:*

$$L(C_\epsilon) = n\big(H(p) + \epsilon'\big) \tag{8}$$

where ϵ' depends on ϵ, n and the cardinality of $\mathcal{X}$.

Proof. Let p be the common distribution of $X_1, , X_n$. Let $A_\epsilon^n(p)$ be the typical set for p corresponding to ϵ and n, henceforth denoted just by A_ϵ^n. Consider an arbitrary bijection $\alpha : A_\epsilon^n \to \big\{1, , |A_\epsilon^n|\big\}$, which assigns to each element $\mathbf{x} \in A_\epsilon^n$ an integer $\alpha(\mathbf{x})$ between 1 and the cardinality of the set A_ϵ^n. Consider another arbitrary bijection $\beta : \mathcal{X}^n \to \big\{1, , |\mathcal{X}|^n\big\}$ which assigns to each element of $\mathbf{x} \in \mathcal{X}^n$ an integer $\beta(\mathbf{x})$ between 1 and the cardinality of the set $\mathcal{X}^n$. For each $\mathbf{x} \in \mathcal{X}^n$, define $C_\epsilon(\mathbf{x})$ as follows: if $\mathbf{x} \in A_\epsilon^n$, then $C_\epsilon(\mathbf{x}) \doteq 0\omega$ (the concatenation of 0 with ω) where ω is the binary representation of the integer $\alpha(\mathbf{x})$; if $\mathbf{x} \notin A_\epsilon^n$, then $C_\epsilon(\mathbf{x}) = 1\omega$ (the concatenation of 1 with ω) where ω is the binary representation of the integer $\beta(\mathbf{x})$. The code C_ϵ is trivially non-singular. Note that for every $\mathbf{x} \notin A_\epsilon^n$, the length $\ell_{C_\epsilon}(\mathbf{x})$ can be bound as follows, where in the fourth step I have recalled that ω is the binary representation of the integer $\beta(\mathbf{x})$ which is smaller than $|\mathcal{X}|^n$.

$$\begin{aligned} \ell_{C_\epsilon}(\mathbf{x}) &= \ell\big(C_\epsilon(\mathbf{x})\big) \\ &= \ell(0\omega) \\ &= 1 + \ell(\omega) \\ &\leq 1 + \big\lceil n \log |\mathcal{X}| \big\rceil \\ &\leq 2 + n \log |\mathcal{X}| \end{aligned} \tag{9}$$

Furthermore, for every $\mathbf{x} \in A_\epsilon^n$, the length $\ell_{C_\epsilon}(\mathbf{x})$ can be bound as follows, where in the fourth step I have recalled that ω is the binary representation of the integer $\alpha(\mathbf{x})$ which is smaller than the cardinality of A_ϵ^n which is in turn smaller than $2^{n(H(p)+\epsilon)}$, as proved above.

$$\begin{aligned} \ell_{C_\epsilon}(\mathbf{x}) &= \ell\big(C_\epsilon(\mathbf{x})\big) \\ &= \ell(1\omega) \\ &= 1 + \ell(w) \\ &\leq 1 + \big\lceil n\big(H(p) + \epsilon\big) \big\rceil \\ &\leq 2 + n\big(H(p) + \epsilon\big) \end{aligned} \tag{10}$$

I can now bound the expected length of the code C_ϵ as follows:

$$\begin{aligned}
L(C_\epsilon) &= \sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x})\ell_{C_\epsilon}(\mathbf{x}) \\
&= \sum_{\mathbf{x}\in A_\epsilon^n} p(\mathbf{x})\ell_{C_\epsilon}(\mathbf{x}) + \sum_{\mathbf{x}\notin A_\epsilon^n} p(\mathbf{x})\ell_{C_\epsilon}(\mathbf{x}) \\
&\overset{(a)}{\leq} \sum_{\mathbf{x}\in A_\epsilon^n} p(\mathbf{x})\Big(2+n\big(H(p)+\epsilon\big)\Big) + \sum_{\mathbf{x}\notin A_\epsilon^n} p(\mathbf{x})\Big(2+n\log|\mathcal{X}|\Big) \\
&= \mathbf{P}\big(A_\epsilon^n\big)\Big(2+n\big(H(p)+\epsilon\big)\Big) + \mathbf{P}\big((A_\epsilon^n)^C\big)\Big(2+n\log|\mathcal{X}|\Big) \\
&\overset{(b)}{\leq} \Big(2+n\big(H(p)+\epsilon\big)\Big) + \epsilon\Big(2+n\log|\mathcal{X}|\Big) \\
&= n\big(H(p)+\epsilon\big)+\epsilon'
\end{aligned}$$

where in step (a) I have used both (9) and (10) and in step (b) I have used the trivial fact that $\mathbf{P}(A_\epsilon^n) \leq 1$ together with the fact that $\mathbf{P}\big((A_\epsilon^n)^C\big) \leq \epsilon$ for n large enough, given that $\mathbf{P}(A_\epsilon^n) \geq 1-\epsilon$, as proven above. □

5 Instantaneous codes and Kraft Inequality

Definition 3. Let C be a code for a random variable X with range $\mathcal{X}$ by means of an alphabet $\mathcal{V}$. The *extension* of C is the function C^* defined a follows:

$$\begin{array}{rccl} C^*: & \mathcal{X}^* & \to & \mathcal{V}^* \\ & x_1 x_n & \mapsto & C^*(x_1 x_n) = C(x_1)C(x_n) \end{array} \tag{11}$$

namely the function which maps any finite-length string x_1x_n of symbols of $\mathcal{X}$ into the string $C(x_1)C(x_n)$ obtained by concatenating in the same order the corresponding codewords. A code C is called *uniquely decidable* if its extension C^* is an injective function. The code C is called a *prefix* or *instantaneous* or *self-punctuating* code if there are no two $x_1, x_2 \in \mathcal{X}$ such that $C(x_1)$ is a prefix of $C(x_2)$.

Osservation 1. For concreteness, let $\mathcal{V} = \{0, 1\}$ and consider the binary tree with infinite height defined as follows:

1. each node has exactly 2 children;
2. the root is labeled with the empty string ϵ;
3. each non-root node is labeled with a word $\omega \in \mathcal{V}^*$;
4. if a node is labeled ω, then its left child is label $\omega 0$ (namely, the concatenation of ω with 0) and its right child is labeled $\omega 1$ (namely, the concatenation of ω with 1).

This infinite binary tree will be called the tree *associated* to $\{0, 1\}^*$. The extension of this construction to arbitrary alphabets $\mathcal{V}$ is of course trivial. Note that the length of a string $\omega \in \mathcal{V}^*$ is the height of the corresponding node in the tree associated with $\mathcal{V}^*$, namely the length of the path from the root to the node labeled with ω. Furthermore, a word ω is a prefix of another word ω' iff ω dominates ω' in the tree associated with $\mathcal{V}^*$. Thus, a code $C : \mathcal{X} \to \{0, 1\}^*$ is instantaneous iff there are no two $x_1, x_2 \in \mathcal{X}$ such that $C(x_1)$ dominates $C(x_2)$ in the tree associated with $\mathcal{V}^*$.

Theorem 4. KRAFT'S INEQUALITY. *[1] Consider a discrete random variable X with finite range $\mathcal{X}$. If $C : \mathcal{X} \to \mathcal{V}^*$ is an instantaneous code for X over an alphabet $\mathcal{V}$, then the following inequality holds:*

$$\sum_{x\in\mathcal{X}} \left(\frac{1}{|\mathcal{V}|}\right)^{\ell_C(x)} \leq 1 \tag{12}$$

[2] Conversely, given $|\mathcal{X}|$ integers l_x with $x \in \mathcal{X}$ such that the above inequality holds, namely:

$$\sum_{x \in \mathcal{X}} \left(\frac{1}{|\mathcal{V}|}\right)^{l_x} \leq 1$$

then there is an instantaneous code $C : \mathcal{X} \to \mathcal{V}^*$ such that $\ell_C(x) = l_x$.

Proof. [1] For concreteness, consider the case $\mathcal{V} = \{0, 1\}$; the extension of the proof to the arbitrary case is trivial. Let $l = \max\{\ell_C(x) \,|\, x \in \mathcal{X}\}$. Consider the tree associated with $\mathcal{V}^*$, as defined in the preceding Observation. For each node x, let $D(x)$ be the set of nodes in the tree which are descendants of x and have height l. Note that

$$|D(x)| = 2^{l - \ell_C(x)} \tag{13}$$

Note furthermore that the following holds, given that the code C is instantaneous:

$$D(x) \cap D(x') = \emptyset \qquad \text{for every } x, x' \in \mathcal{X} \tag{14}$$

Hence:

$$\begin{aligned} 2^l &\geq \left| \bigcup_{x \in \mathcal{X}} D(x) \right| \\ &= \sum_{x \in \mathcal{X}} |D(x)| \\ &= \sum_{x \in \mathcal{X}} 2^{l - \ell(x)} \end{aligned}$$

where: in the first step, I have noted that there are 2^l nodes of height l and I have recalled that each $D(x)$ is by definition a set of nodes of height l; in the second step, I have used (14) and in the third step I have sued (13). The claim follows by dividing both sides by 2^l.

[2] Pick a node ω_1 of height l_1 in the tree associated with $\{0, 1\}^*$, let $C(x_1) = \omega_1$ and remove all the descendants of ω_1 from the tree; pick a node ω_2 of height l_2 in the tree associated with $\{0, 1\}^*$, let $C(x_2) = \omega_2$ and remove all the descendants of ω_2 from the tree; and so on. In this way, we build a prefix code with the assigned code lengths. □

6 Compression via Kraft's Inequality

Theorem 5. *Consider a discrete random variable X with finite range $\mathcal{X}$. The minimum expected length of an instantaneous code for X is $H(p)$.*

Proof. Assume that $\mathcal{X} = \{1, , M\}$. By virtue of Kraft's Inequality, the instantaneous code which achieves the minimum expected length is the code whose codeword lengths $l_1, , l_M$ solve the following constrained optimization problem:

$$\begin{aligned} &\text{minimize:} && \textstyle\sum_{i=1}^M p_i l_i \\ &\text{subject to:} && \textstyle\sum_{i=1}^M 2^{-l_i} \leq 1, \\ & && l_i \geq 0 \text{ for } i = 1, , M \end{aligned}$$

We can solve this problem by means of the method of Lagrange multipliers. The corresponding Lagrangian is as follows:

$$\Lambda(l, \lambda) = \sum_{i=1}^M p_i l_i + \lambda \left(\sum_{i=1}^M 2^{-l_i} - 1 \right) \tag{15}$$

and its derivatives are as follows:

$$\frac{\partial \Lambda(l, \lambda)}{\partial l_i} = p_i - \lambda 2^{-l_i} \ln 2 \tag{16}$$

Thus, the condition $\frac{\partial \Lambda(l,\lambda)}{\partial l_i} = 0$ holds iff the following holds:

$$2^{-l_i} = \frac{p_i}{\lambda \ln 2} \tag{17}$$

By imposing that the constrains are satisfied, I get the following:

$$1 = \sum_{i=1}^{M} 2^{-l_i} = \sum_{i=1}^{M} \frac{p_i}{\lambda \ln 2} = \frac{1}{\lambda \ln 2} \tag{18}$$

from which I derive that $\lambda = \frac{1}{\ln 2}$. By replacing this expression for λ in (17), I conclude that $p_i = 2^{-l_i}$ and thus:

$$l_i = -\log p_i = \log \frac{1}{p_i} \tag{19}$$

Thus, the optimal code C_{opt} has code lengths $l_i = \log \frac{1}{p_i}$ and its expected length is:

$$\ell(C_{\text{opt}}) = \sum_{i=1}^{M} p_i l_i = \sum_{i=1}^{M} p_i \log \frac{1}{p_i} = H(p) \tag{20}$$

namely, the entropy of X. $\square$