Mathematical methods in communication June 31st, 2010

Lecture 9

Lecturer: Haim Permuter *Scribe: Imri Zviely and Itai Lishner*

I. Maximum Entropy, Channel coding and Side information

In the nature many phenomenons may be explained by the maximum entropy principle. For instance, the velocity of the particles in the air for a given temperature is distributed according to Gaussian distribution. The temperature is proportional to the kinetic energy of the particles. Given a specific temperature, namely, a mean square constraint on the velocity, we obtain that a Gaussian distribution maximizes the entropy. In this lecture we would learn how to calculate the maximum entropy, and more important, understand why this is the case in nature.

Consider the following problem: we would like to maximize the entropy $H(X)$ over $p(x)$ such that the following constraints will hold

1) $p(x) \geq 0, \quad \forall x \in \mathcal{X}$.
2) $\sum_{\in \mathcal{X}} p(x) = 1$.
3) $\sum_{x \in \mathcal{X}} p(x) r_i(x) = \alpha_i, \quad i = (1, ..., m)$,

where $r_i(x)$ are functions of x. Note that the constraint $\sum_{x \in \mathcal{X}} p(x) r_i(x) = \alpha_i$ is equivalent to $E[r_i(X)] = \alpha_i$.

Theorem 1 The optimal pmf $p(x)$ for the optimization problem given above is

$$p^*(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)} \tag{1}$$

where λ_i, $i = 0, 1, 2, ..., m$ is the one that satisfies

$$\begin{aligned} \sum_x p^*(x) &= 1 \\ \sum_x p^*(x) r_i(x) &= \alpha_i. \end{aligned} \tag{2}$$

Q: Is this problem a convex optimization problem ?

A: Yes.

We learned that a convex optimization problem is of the form

$$\min_x \quad f_0(x) \tag{3}$$

$$s.t. \quad f_i(x) \leq 0, 1 \leq i \leq m \ , \tag{4}$$

$$h_j(x) = 0, 1 \leq j \leq l \tag{5}$$

where $f_i,\ i = 0, 1, 2, ...m$ are convex functions and h_j are affine functions.

In our case we have the problem

$$\max_{P_X} \quad H(P_X) \tag{6}$$

$$s.t. \quad P_X(x) \geq 0,\ \forall x \in \mathcal{X}, \tag{7}$$

$$\sum_{x\in\mathcal{X}} P_X(x) = 1, \tag{8}$$

$$\sum_{x\in\mathcal{X}} P_X(x) r_i(x) = \alpha_i i = (1, ..., m). \tag{9}$$

Proof of Theorem 1

We solve the optimization problem using the Dual lagrange principle. We first dismiss the constraint that $p(x) \geq 0$, however since we will obtain that the optimal solution $p^*(x)$ satisfies this condition, then $p^*(x)$ is optimal also with this constraint.

We form the Lagrange dual functional:

$$J(p(x), \overline{\lambda}) \quad = \quad -\sum p(x)\log(p(x)) + \lambda_0(\sum p(x) - 1) + \sum_{i=1}^{m} \lambda_i (\sum_x p(x) r_i(x) - \alpha_i) \tag{10}$$

And differentiate with respect to $p(x)$:

$$\frac{\partial J(p(x), \overline{\lambda})}{\partial p(x)} \quad = \quad -\log(p(x)) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) = 0 \tag{11}$$

Hence we obtained that

$$p(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)} \tag{12}$$

Now we need to find λ_0 and $\lambda_i,\ i = 1, 2, ..., m$ such that

$$\begin{aligned} \sum_x p(x) &= 1 \\ \sum_x p(x) r_i(x) &= \alpha_i. \end{aligned} \tag{13}$$

For the continuous alphabet where we want to maximize the differential entropy over differential distributions $f(x)$ that satisfies the constraints $E[r_i(X)] = \alpha_i,\ i = 1, 2, ..., m$ we obtain

$$f(x) \quad = \quad e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)} \tag{14}$$

such that

$$\int_x f(x)dx \quad = \quad 1 \tag{15}$$

$$\int_x f(x) r_i(x) dx \quad = \quad \alpha_i. \tag{16}$$

$$\tag{17}$$

Example 1 (dice) Consider a dice with six faces $[1, 2, ..., 6]$. What would be the probability of the dice that maximize the entropy?

$$\left.\begin{array}{l} p(x) = e^{\lambda_0 - 1} \\ \sum_x p(x) = 1 \end{array}\right\} \quad \Rightarrow \quad p^*(x) = \frac{1}{6} \tag{18}$$

Note that we obtain that $p(x)$ does not depend on x, namely is constant. Therefore,a uniform distribution maximizes the entropy.

Example 2 : $X \in [-\infty, \infty]$ Let the constraints be $E[X] = 0$ and $E[X^2] = \sigma^2$. Then the form of the maximizing distribution is

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2} \tag{19}$$

To find the appropriate constants, we first recognize that this distribution has the same form as normal distribution. Hence, the density that satisfies the constraints and also maximizes the entropy is the $\mathcal{N}(0, \sigma^2)$ distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma 2}} e^{-\frac{x^2}{2\sigma^2}} \tag{20}$$

Alternative proof to Theorem 1 Let $P_X(x) = e^{\lambda_0 - 1 + \sum i=1^m \lambda_i r_i(x)}$, and let $Q_X(x)$ be any distribution that satisfies the conditions $E[r_i(x)] = \alpha_i$.

We need to show that $H(P_X) \geq H(Q_X)$. Consider,

$$\begin{aligned} D(Q\|P) &= \sum_x Q(x) log \frac{Q(x)}{P(x)} && (21) \\ &= \sum_x Q(x) log(Q(x)) - \sum_x Q(x) log(P(x)) && (22) \\ &= -H(Q) - \sum_x Q(x)[\lambda_0 - 1 + \sum_{i=1}^m \lambda_i r_i(x)] && (23) \\ &\overset{(a)}{=} -H(Q) - \sum_x P(x)[\lambda_0 - 1 + \sum_{i=1}^m \lambda_i r_i(x)] && (24) \\ &\overset{(b)}{=} -H(Q) + H(P), && (25) \\ & && (26) \end{aligned}$$

and since $D(P||Q) \geq 0$ we obtained that $H(P) \geq H(Q)$.

Example 3 : Dice x , $\mathcal{X} = [1, 2, ..., 6]$

Suppose that n dice are thrown on the table and we are told that the sum of the results is $n\alpha$. What is the probability p(x)?

It is easy to show that $E[x] = \alpha$.

We define $n_1, n_2, ..., n_6$, when n_i represents the number of throws where $x = i$.

Number of sequences with $(n_1, n_2, ..., n_6)$ is equal to:

$$\begin{pmatrix} n \\ n_1 \quad n_2 \quad ... \quad n_6 \end{pmatrix} = \frac{n!}{n_1!\ n_2!\ ...\ n_6!} \tag{27}$$

In order to find the most probable state we wish to maximize $\frac{n!}{n_1!\ n_2!\ ...\ n_6!}$ under the constraint:

$$\sum_{i=1}^{6} \frac{n_i}{n} \cdot i = \alpha \qquad ; \qquad E[x] = \alpha \tag{28}$$

For large values of n using Stirling's approximation, $n! \approx (\frac{n}{e})^n$, we define that:

$$N = \frac{n!}{n_1!\ n_2!\ ...\ n_6!} \approx \frac{(\frac{n}{e})^n}{(\frac{n_1}{e})^{n_1} \quad ... \quad (\frac{n_6}{e})^{n_6}} \tag{29}$$

$$log(N) = nlog(n) - n_1 log(n_1) - ... - n_6 log(n_6) \tag{30}$$

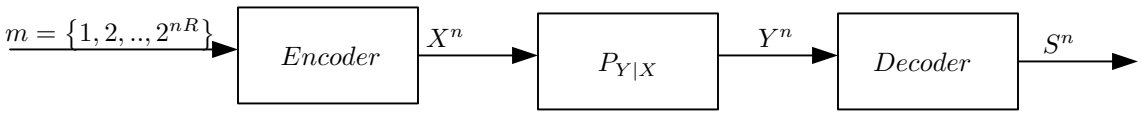
$$n = n_1 + n_2 + ... + n_6 \tag{31}$$

$$log(N) = -\sum_i n_i log \frac{n_i}{n} \tag{32}$$

$$N = 2^{-\sum_i n_i log \frac{n_i}{n}} = 2^{-n \sum_i \frac{n_i}{n} log \frac{n_i}{n}} = 2^{nH(\frac{n_1}{n}, \frac{n_2}{n}, ..., \frac{n_6}{n})} \tag{33}$$

II. Channel Coding With Side Information

Channel coding:



$$C = \max_{p(x)} I(X;Y) \tag{34}$$

Fig. 1. Communication system

Channel coding with Side Information:

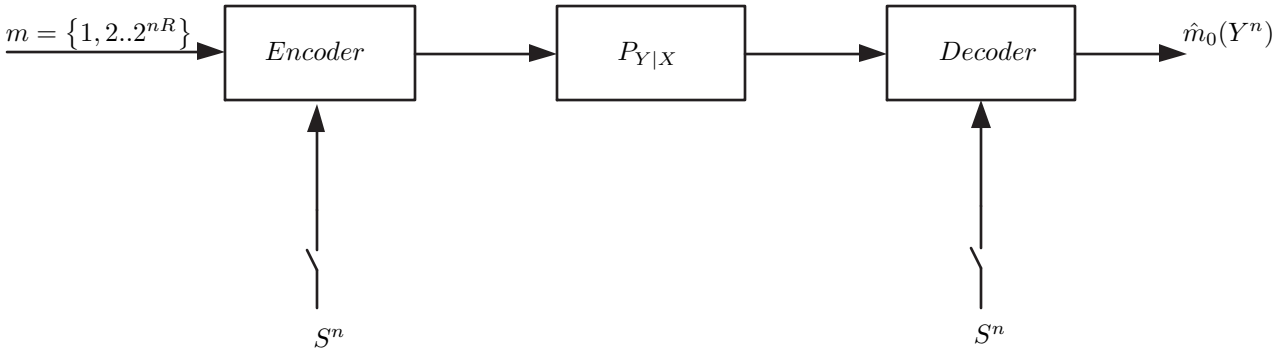


Fig. 2. Communication system with Side Information

$$P(y_i|x^i, s^i, y^{i-1}) = P(y_i|x_i, s_i) \tag{35}$$

$$P(y^n|x^n, s^n) = \prod_{i=1}^{n} P(y_i|x_i, s_i) \quad (without feedback) \tag{36}$$

Case I : state information known only to the Decoder:

$$C = \max_{p(x)} I(X;Y|S) \overset{(a)}{=} I(X;Y,S) \tag{37}$$

$a)$ $p(s)p(x)p(y|x,s) = p(x,y,s)$

We define the problem:

$$Encoder \quad : \quad f : \{1, 2, ..., 2^{nR}\} \ \rightarrow \ X^n$$

$$Decoder \quad : \quad g : Y^n, S^n \ \rightarrow \ \{1, 2, ..., 2^{nR}\}$$

$$S^n \sim p(s) \quad i.i.d$$

For memoryless channel:

$$P(y_i, s_i|x^i, y^{i-1}, s^{i-1}) = P(s_i)P(y_i|x^i, y^{i-1}, s^i =$$

$$= P(s_i)P(y_i|x_i, s_i) = P(y_i, s_i|x_i)$$

Case II : state information known to both Encoder and Decoder:

$$C = max_{p(x|s)} I(X;Y|S) \tag{38}$$

We define the problem:

$$\begin{aligned} Encoder &: f : \{1, 2, ..., 2^{nR}\} x S^n \rightarrow X^n \\ Decoder &: g : Y^n, S^n \rightarrow \{1, 2, ..., 2^{nR}\} \end{aligned}$$

proof :

$$C = \max_{p(x|s)} \sum_s p(s) I(X; Y|S = s) \tag{39}$$

We will split the message $\{1, 2, ..., 2^{nR}\}$: $\{1, 2, ..., 2^{nR_0}\}$ x $\{1, 2, ..., 2^{nR_1}\}$

$$R_0 = I(X; Y|S = 0) \tag{40}$$

$$R_1 = I(X; Y|S = 1) \tag{41}$$

Doing that we split the channel into two separate channels:

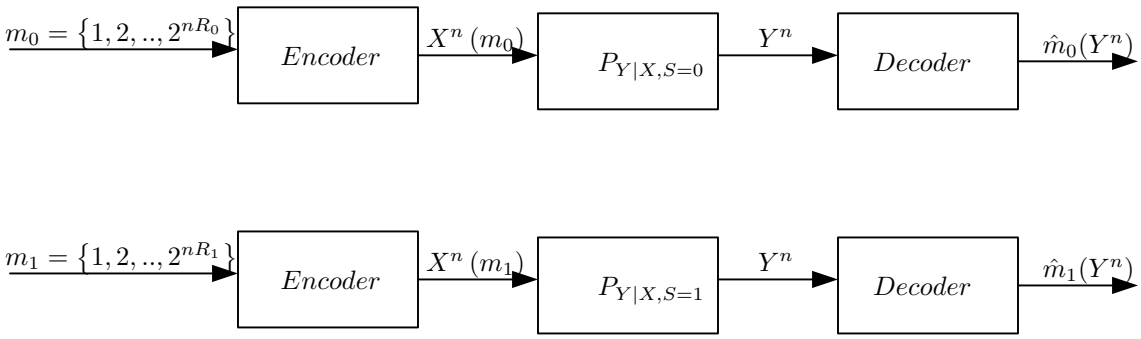


Fig. 3. Splitted Channel

such that:

$$\begin{aligned} R_0 &= n \cdot p(S = 0) I(X; Y|S = 0) \\ BlockSize(S = 0) &= n \cdot p(S = 0) \\ R_1 &= n \cdot (S = 1) I(X; Y|S = 1) \\ BlockSize(S = 1) &= n \cdot p(S = 1) \end{aligned}$$

Calculating the total rate of the channel:

$$R = \frac{1}{n}[np(S = 0) I(X; Y|S = 0) + np(S = 1) I(X; Y|S = 1)] \overset{(a)}{=} I(X; Y|S) \tag{42}$$

(a) Law of large numbers

Converse :

$$
\begin{aligned}
nR &= H(M) \qquad (43)\\
&\overset{(a)}{=} H(M|S^n)\\
&= H(M|S^n) - H(M|S^n, Y^n) + H(M|S^n, Y^n)\\
&\overset{(b)}{\leq} I(M; Y^n|S^n) + n\epsilon_n\\
&= I(M, X^n(M, S^n); Y^n|S^n) + n\epsilon_n\\
&= H(Y^n|S^n) - H(Y^n|S^n, X^n, M) + n\epsilon_n\\
&\overset{(c)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1}, S^n) - H(Y_i|Y^{i-1}, S^n, X^n, M) + n\epsilon_n\\
&\leq \sum_{i=1}^{n} H(Y_i|S_i) - H(Y_i|S_i, X_i) + n\epsilon_n\\
&= \sum_{i=1}^{n} I(Y_i; X_i|S_i) + n\epsilon_n\\
&\leq [\max_{p(x|s)} I(Y; X|S) + \epsilon_n]n
\end{aligned}
$$

(a) $S^n \perp M$

(b) Fano inequality

(c) Chain rule