Indian Institute of Technology Bombay

Department of Electrical Engineering

Handout 5 EE 708 Information Theory and Coding

Tutorial 2 Feb 3, 2018

Question 1) Let us extend our definition of entropy to include all discrete random variables on Ω as

$$H(X) = \sum_{x \in \Omega} p(x) \log_e \frac{1}{p(x)},$$

where we removed the constraint on the cardinality of Ω. Among all non-negative random variables X with mean at most μ, find the maximum value of $H(X)$.

Solution: A standard way to solve such problems is to use Lagrange multipliers to factor in the constraint. However, as we mentioned in class, there are simpler techniques as well. We will describe both methods, one after the other.

Lagrange Optimization: Since many of you may be unaware of the Lagrange techniques, lemme start with a simple one dimensional example. Suppose you have to maximize $g(x)$ in the interval $-\infty < x < \infty$. We can find the extremas by solving for the first derivative $g'(x) = 0$. Of course, we assume differentiability etc, and do not worry too much about existences of the quantities we compute. If there is only one maxima, we can pick it from the solutions by using the second derivative.

Now, imagine we need to maximize $g(x)$ for $-\infty < x \leq a$ for some finite value of a. It is not difficult to see that

$$\max_{-\infty < x \leq a} g(x) = \max_{x \in \mathbb{R}} g(x) + \min_{\lambda < 0} \lambda(x - a).$$

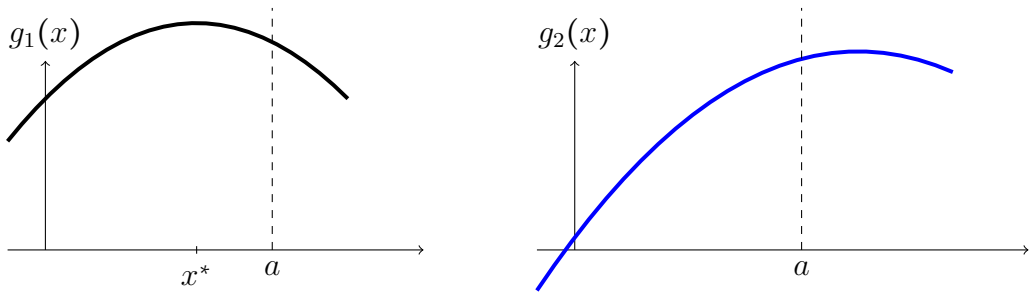
Okay, it is a bit of a juggle, but no magic. Whenever $x > a$, the right hand side above returns the value $-\infty$, since lambda can be taken to be a negative value with arbitrary high magnitude. So the maximum cannot happen when $x > a$. On the other hand, with $x < a$, the *optimal* value of λ is 0, as negative values of λ will only hamper the minimization since $(x - a)$ is already negative. That leaves us with the specific case of $x = a$. Notice that when $x = a$, the value of λ is something we can choose, as $\lambda(x - a) = 0$, no matter what λ we pick. For a given $\lambda \leq 0$, suppose we optimize the cost function

$$J(x, \lambda) = \max_{x \in \mathbb{R}} g(x) + \lambda(x - a), \tag{1}$$

using our standard approach, i.e. solve for

$$J'(x, \lambda) = g'(x) + \lambda = 0. \tag{2}$$

The pictures below illustrate the meaning of this step, where $\lambda = -g'(x)$ is negative only for the graph on the right.



Putting it altogether, we can solve for (1) for an appropriate Lagrange multiplier $\lambda < 0$, and figure out the x^*, if the latter exists. The beautiful theory of convex optimization will tell you about the guarantees of a solution, and also the computational aspects. If there are more than one variable, we can check the derivative w.r.t each variable. Furthermorei, each new constraint can be given a separate Lagrange multiplier as well.

For the problem at hand, by defining $\bar{p} = \{p(x), x \in \Omega\}$, the cost function becomes

$$J(\bar{p}, \lambda, \nu) = \sum_{x\in\Omega} p(x) \log \frac{1}{p(x)} + \lambda\Big(\sum_{x\in\Omega} xp(x) - \mu\Big) - \nu\Big(\sum_{x\in\Omega} p(x) - 1\Big). \tag{3}$$

We take $\mu > 0$ to avoid trivialities. Just for convenience, I will also take $\nu = 0$ and write the cost as $J(\bar{p}, \lambda)$. A constraint less should only lead to higher objectives here. Differentiating $J(\bar{p}, \lambda)$ w.r.t to the variable $p(x)$ (for some x) and equating to 0, we get

$$p(x)\frac{-p(x)}{p^2(x)} + \log \frac{1}{p(x)} + \lambda x = 0. \tag{4}$$

The $p(x)$ which solves it is of the form

$$p(x) = \exp(\lambda x - 1).$$

In order to make it a probability distribution we can take

$$q(x) = \frac{\exp(\lambda x - 1)}{\sum_x \exp(\lambda x - 1)} = [1 - \exp(\lambda)] \exp(\lambda x). \tag{5}$$

The above step also says why we set $\nu = 0$ in (3), as we could explicitly account for the constraint. We are done if we successfully figure out a meaninful $\lambda < 0$. Using the fact that $\mathbb{E}X \leq \mu$,

$$\mathbb{E}\big[X\big] = [1 - \exp(\lambda)] \sum_{x\in\Omega} x \exp(\lambda x) = \frac{\exp(\lambda)}{1 - \exp(\lambda)} \leq \mu. \tag{6}$$

At equality above, we get $\lambda = \log_e \frac{\mu}{1+\mu}$, and consequently the cost function becomes

$$J(p^*, \lambda) = \frac{1}{1+\mu} + \mu \log \frac{\mu}{1+\mu}.$$

The RHS is an increasing function of $\mu \geq 0$, this is easily seen by computing its first derivative w.r.t μ. So it is indeed optimal to consider $\mathbb{E}[X] = \mu$ while computing the maximal entropy. Substituting $\lambda = \log \frac{\mu}{1+\mu}$ in (5), the optimal distribution p^* can be evaluated as

$$p^*(x) = \left(\frac{\mu}{1+\mu}\right)^x \left(\frac{1}{1+\mu}\right), \; x \in \{0, 1, \ldots\}, \tag{7}$$

and the maximal entropy is

$$H(X) = \frac{1}{1+\mu} + \mu \log \frac{\mu}{1+\mu}.$$

Recall from your probability classes that (7) is nothing buy a **geometric distribution**. To illustrate further, take $\frac{1}{1+\mu}$ as the probability of HEAD in tossing a coin. Let Y be a random variable which specifies the number of IID tosses to obtain a HEAD for the first time. Then $Y - 1$ is the entropy maximizing random variable on non-negative integers, under a mean constraint of μ.

Alternate Method: We do not claim any superiority of the second method, described now, other than a certain elegance. The recipe builds on a certain guess for the optimal distribution. Let

$$q(x) = [1 - \exp(\alpha)] \exp(\alpha x),$$

for some negative parameter α. Notice that

$$\sum_{x \in \Omega} q(x) \log q(x) = \log(1 - \exp(\alpha)) + \alpha \mathbb{E}[X] = \sum_{x \in \Omega} p(x) \log q(x).$$

From the theorem given in class, now $q(x)$ is the entropy maximizing distribution, for an appropriate value of α. We can find $\alpha = \log \frac{\mu}{1+\mu}$ from $\mathbb{E}[X] = \mu$, similar to (6).

Question 2) Let $\mathcal{X} = \{1, 2, \ldots, 10\}$ be the set of possible values of X. Show that among all probability distributions with the first moment at most 5, the one which maximizes $H(X)$ has the form $P(X = x) \propto \exp(-\lambda x)$. Suggest how you will find λ. How will your answer change if $\mathcal{X} = \{-5, -4, \ldots, 4, 5\}$.
Solution: Similar to the last problem, we can take

$$q(x) = \frac{\exp(\lambda x)}{\sum_{x \in \mathcal{X}} \exp(\lambda x)}, \tag{8}$$

to get $\sum_{x \in \mathcal{X}} q(x) \log q(x) = \sum_{x \in \mathcal{X}} p(x) \log p(x)$, for any $p(x)$ such that $\sum_{x \in \mathcal{X}} x p(x) = \sum_{x \in \mathcal{X}} x q(x)$. In other words, $q(x)$ is entropy maximizing in its *class*. The parameter λ can be chosen to meet the mean constraint.

It was observed in class that if a mean constraint of $\mathbb{E}[X] \leq 5.5$ is imposed, then the uniform distribution itself maximizes the entropy. This does not contradict our answer in (8), since

$$\sum_{1 \leq x \leq 10} x \frac{\exp(\lambda x)}{\sum_{x \in \mathcal{X}} \exp(\lambda x)} = 5.5,$$

will imply that $\lambda = 0$ is the only solution, leading to the uniform distribution as the optimal distribution.

Question 3) For given distributions $p_1(x)$ and $p_2(x)$, define $p_\lambda(x) := \lambda p_1(x) + (1 - \lambda) p_2(x)$ for some $0 \leq \lambda \leq 1$. Show that

$$H(X_\lambda) \geq \lambda H(X_1) + (1 - \lambda) H(X_2),$$

where $X_j \sim p_j(x)$ for $j \in \{1, 2, \lambda\}$.

Solution:

$$\begin{aligned}
&\lambda \sum_x p_1(x) \log \frac{1}{p_1(x)} + (1 - \lambda) \sum_x p_2(x) \log \frac{1}{p_2(x)} \\
&\quad - \sum_x \big[\lambda p_1(x) + (1 - \lambda) p_2(x)\big] \frac{1}{\lambda p_1(x) + (1 - \lambda) p_2(x)} \\
&\qquad = \lambda \sum_x p_1(x) \log \frac{\lambda p_1(x) + (1 - \lambda) p_2(x)}{p_1(x)} \\
&\qquad\quad + (1 - \lambda) \sum_x p_2(x) \log \frac{\lambda p_1(x) + (1 - \lambda) p_2(x)}{p_2(x)} \\
&\qquad = -\Big[\lambda D\Big(p_1 \| \lambda p_1 + (1 - \lambda) p_2\Big) + (1 - \lambda) D\Big(p_2 \| \lambda p_1 + (1 - \lambda) p_2\Big)\Big] \\
&\qquad \leq 0,
\end{aligned}$$

where the last step used $D(p\|q) \geq 0$ (or $log_e(x) \leq x - 1$). Thus $H(X_1) + (1-\lambda)H(X_2) - H(X_\lambda) \leq 0$, completing the proof.

Question 4) Consider a Markov chain $X_n, n \geq 1$ on $\Omega = \{a, b, c, d\}$. The transition probabilities are given by the matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \alpha & \beta & \gamma & \delta \\ \delta & \alpha & \beta & \gamma \\ \gamma & \delta & \alpha & \beta \end{bmatrix},$$

where $\alpha, \beta, \gamma, \delta$ are positive numbers with their sum as unity. Notice that $\mathbf{P}_{ij}$ denotes the probability $P(X_{n+1} = j | X_n = i)$. In order to encode a long sequence of values taken by this Markov chain, what do you think is the minimum average length of any prefix free code that can be used. Can you suggest an encoding scheme.

Solution: Imagine a scheme where you encode k consecutive values of the chain using a Huffman code, and then proceed to the next k symbols, and so on. We know that

$$\mathbb{E}[L_k] \leq H(X_1, \ldots, X_k) + 1,$$

number of bits suffices on the averaga to encode $X_1, \ldots X_k$. In other words, the average number of bits per source symbol L_{avg} is

$$L_{avg} = \frac{1}{k}\mathbb{E}[L_k] = \frac{1}{k}H(X_1, \ldots, X_k) + \frac{1}{k}.$$

We know that $p(x_{n+1}|x_n, x_{n-1}, \ldots, x_1) = p(x_{n+1}|x_n)$. Computing the joint entropy,

$$\begin{aligned} H(X_1, \ldots, X_k) &= H(X_1) + \sum_{i=2}^{k} H(X_i | X_{i-1}, \ldots, X_1) \\ &= H(X_1) + \sum_{i=2}^{k} H(X_i | X_{i-1}) \\ &= H(X_1) + (k-1)H(X_2|X_1). \end{aligned}$$

In the last step, we assumed to operate on the steady state distribution of the Markov chain, that the time indexes are irrelevant. Since $H(X_1) \leq \log_2 |\mathcal{X}| = 2$, the average length per symbols becomes

$$L_{avg} \leq \frac{k-1}{k} H(X_2|X_1) + \frac{3}{k},$$

which becomes close to $H(X_2|X_1)$ as k becomes large. The general principle of cutting a sequence of dependent random variables to large enough chunks, and dealing with them as *essentially* independent segments is useful in other contexts as well.

Question 5) Consider an IID source on $\mathcal{X} = \{A, B, C\}$, with probabilities $(0.7, 0.2, 0.1)$. A long sequence of the source symbols is to be stored in the memory after proper encoding. Suppose we can store only octal numbers (i.e. $0, 1, \ldots, 7$, or the memory consists of three bit registers connected to a bus, that you can read and write together). The punctuation requirement is that every time you read a memory location, all source symbols encoded till that time should be revealed. Suggest a scheme and determine its compression efficiency, in number of bits per source symbol.

Solution: The essential idea is that of Tunstall Coding. We have 8 possible labels using 3 bits or an octal value. Let us take the labels as $\{0, 1, 2, 3, 4, 5, 6, 7\}$.

The following is a possible map.

A	5
B	6
C	7

The average encoding efficiency is 1 label per source symbol. We can pack more sources for one label, for example by the following map,

AA	3
AB	4
AC	5
B	6
C	7

we can get $\frac{0.7}{2} + 0.2 + 0.1 = 0.65$ labels per source symbol. Essentially, whenever a new A happens, 2 symbols are sent within one label, or only 0.5 labels per symbol is required whenever A is observed next. Building on this, let us consider the following two possibilities.

AA	1
AB	2
AC	3
BA	4
BB	5
BC	6
C	7

AAA	1
AAB	2
AAC	2
AB	4
AC	5
B	6
C	7

Which one do you think is more efficient? Now that will tell you that the highest probability symbol at any stage is further split if sufficiently many labels are remaining, in general.