HW05 STAT W4400

Ethan Grant uni: erg2145

April 28, 2016

1. (a)

$$D = 1, 2, 4, 8$$

$$H(x) = -(\sum_{i \epsilon D} p_i * log(p_i) + L(p_1, ..., p_d, \lambda_0, \lambda_1))$$

$$= -(\sum p_i * log(p_i) + \lambda_0 * p_i + \lambda_1 * (\sum p_i - 1)$$

$$\frac{dL}{dp_i} = \frac{-p_i}{p_i} - log(p_i) + \lambda_1 = 0$$

$$p_i* = exp(\lambda_1 - 1)$$

The value of p_i does not depend on i because λ_1 does not depend on i and thus p_i must have the same value for all i meaning that the maximum entroy solution is the uniform distribution for each distribution

(b) To have minimum entropy all the mass for each of the distributions would be at a particular point as described by the dirac. So every distribution equals its dirac (represented by delta) $p(X = i) = \delta_i$ for all i in the finite set 1,2,4,8

(c) There are 2 options for each x_i in the sequence. Thus there are 2^n total possible sequences. The result from a tells us that the uniform distribution will have the maximum entropy. From the slides we know that $H(U_d) = log_2(2^n) = n * log_2(2^n) = n$ where U_d is the uniform distribution.

(d) We must specify both the start and the transition matrix. Every element in the matrix (which specifices the probability of moving from one value to another) will be the same because it is the uniform distribution. Since there are 2 rows and each column must sum to 1 you know the transition matrix has .5 at each category. Similarly the start must sum to 1 and have equal probability for all elements and since there are 2 elements it is as it is represented below

$$\text{P} = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}$$

$$P_{init} = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$$

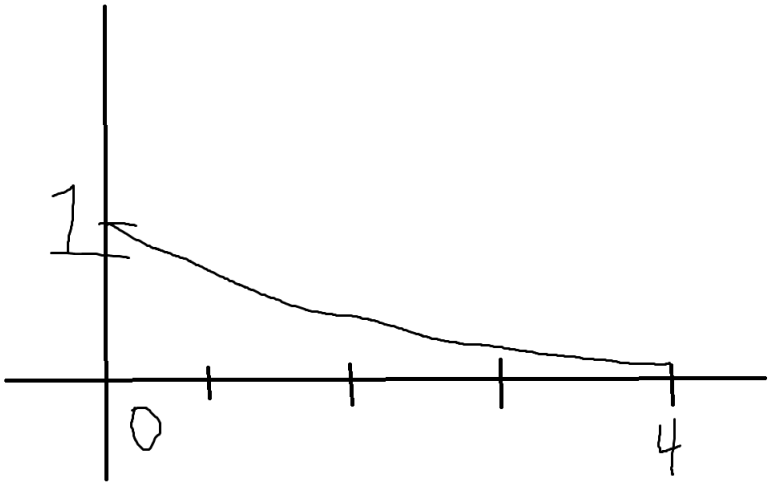
(e) Since we have already found the maximum entropy solution and this solution is not the maximum entropy solution it must have lower entropy than the chain above.

(f) P = $\begin{bmatrix} .6 & .3 \\ .4 & .7 \end{bmatrix}$
$det(\begin{bmatrix} .6-\lambda & .3 \\ .4 & .7-\lambda \end{bmatrix}) = (.6-\lambda)(.7-\lambda) - .12 = 0$
$\lambda_1 = .3 \lambda_2 = 1$
B/C there is a lambda with a value 1 we know this chain converges
$E_1 = ker(\begin{bmatrix} -.4 & .3 \\ .4 & -.3 \end{bmatrix}) = ker(\begin{bmatrix} 1 & .75 \\ 0 & 0 \end{bmatrix}) = \begin{bmatrix} .75 \\ 1 \end{bmatrix}$
normalize this eigenvector $\frac{\begin{bmatrix} .75 \\ 1 \end{bmatrix}}{\sqrt{.75^2+1}} = \begin{bmatrix} .6 \\ .8 \end{bmatrix}$
Want the columns to sum to 1 so need to do that while maintaining the ratio of $\begin{bmatrix} .6 \\ .8 \end{bmatrix}$ note .6 = 3*.2 and .8 = 4*.2 thus can transform matrix to:
$p_{eq} = \begin{bmatrix} \frac{3}{7} \\ \frac{4}{7} \end{bmatrix}$

(g) By definition $lim_{n->\infty} p_n = p^n * p_{eq} = p_{eq}$ so $lim_{n->\infty}(Pr(x=1)) = 4/7$ and $lim_{n->\infty}(Pr(x=0)) = 3/7$

(h) Important info: $H[X_i|H_{i-1}] = H[X_2|H_1]$ based on the Markov properties since you are starting from the equilibirum vector. Thus you only need to calculate $H[X_2|H_1]$ which will occur n-1 times and $H[X_1]$ to deterimine the entropy of the whole chain
Also note all logs are base 2 as is customary:
$H(X_1) = -4/7 * log(4/7) - 3/7 * log(3/7) = .9852$
Some probabilities that will be used in the following calculation where P(x,y) = $P(x_2, x_1)$:
$P(1,0) = .3 * 3/7$
$P(1,1) = .7 * 4/7$
$P(0,0) = .6 * 3/7$
$P(0,1) = .4 * 4/7$
$H[X_2|H_1] = -(.4*4/7*log(.7) + .3*4/7*log(.3) + .6*3/7*log(.6) + .4*3/7*log(.4) = .91972$
Total Entropy = $.9852 + (n-1) * .91972$

2. (a) Graph: $e^0 = 1 e^{-1} = 1/e ... e^{-4} = 1/(e^4)$



(b) There is no visual representation of the likelihood of an individual point on a pdf. the likelihood is the product of the exponential at the points 1 , 2, 4

(c) The higher rate would decrease the likelihood for the toy data set. Observe that p(1,2) = $2/e^2$ ¡ 1/e , p(2,2)=$2/e^4$ ¡ $1/e^2$, p(4,2)=$4/e^8$ ¡ $1/e^4$. Since the individual likelihoods are all smaller their product is smaller as well and since the product is the definition of likelihood the likelihood declines

(d)

$$q(\theta|x_1...x_n) = \frac{\Pi^n(\theta * e^{-\theta * x_i}) * \frac{\theta^{\alpha_0 - 1} * \beta^{\alpha_0} * e^{-\beta_0 * \theta}}{\Gamma(\alpha_0)}}{p(x_1, ..., x_n)}$$

$$aside : c = \frac{1}{\Gamma(\alpha_0) * p(x_1...x_n)}$$

$$= c * \theta^n * e^{-\theta * \sum(x_i)} * \theta^{\alpha_0 - 1} * \beta^{\alpha_0} * e^{-\beta_0 * \theta}$$

$$= c * \theta^{\alpha_0 + n - 1} * e^{-\theta * (\sum(x_i) + \beta_0}$$

$$= gamma(\theta, \alpha_0 + n, \sum x_i + \beta_0)$$

(e) You can define the value from 1 to n (using all data points) easily based on the definitions provided:

$$\Pi(\theta|x_1, ..., x_n) = \frac{\Pi_i^n p(x_i|\theta) * q(\theta)}{\Pi_1^n p(x_i)}$$

You can then seperate out the nth calculation because you are just multiplying leaving you with the posterior for n-1 calcuations times

the nth observation

$$\Pi(\theta|x_1,...,x_n) = \frac{p(x_n|\theta)}{p(x_n)} \frac{\Pi_i^{n-1} p(x_i|\theta) * q(\theta)}{\Pi_1^{n-1} p(x_i)}$$

thus we can see that prior can now be defined as:

$$q(\theta) = \Pi(\theta|x_1,...,x_{n-1}) = \frac{\Pi_i^{n-1} p(x_i|\theta) * q(\theta)}{\Pi_1^{n-1} p(x_i)}$$

(f) This is based on the answer to problem 2.d. The sum of new x's is merely summing x_n because it is the only new point and thus all that is added to $beta_{n-1}$ and there is only 1 to add to α_{n-1}.

$$g(\theta|x_n) = Gamma(\theta, \alpha_{n-1} + 1, \beta_{n-1} + x_n)$$

(g) As n increases the variance will decrease and the function will have a higher peak that is centered more tightly around the value of theta
The gray line with the highest peak has an n =256
the purple line has n = 16
the blue line has n = 8
the black line has n =4