Mid-term solutions
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2015

1.a) For the exponential distribution

$$\begin{aligned} h[\mathcal{D}] &= -\frac{1}{n}\sum_i(-\lambda x_i + \log\lambda) \\ &= \lambda \times \frac{1}{n}\sum_i x_i - \log\lambda \end{aligned}$$

Hence, for large n

$$\begin{aligned} h[\mathcal{D}] &\approx \lambda E_X[X] - \log\lambda \\ &= 1 - \log\lambda = H[X]. \end{aligned}$$

b)

$$\begin{aligned} & E_{X,Y}[\{f(X) - E_X[f(X)]\}\{f(Y) - E_Y[f(Y)]\}] \\ &= \int P_{X,Y}(x,y)\{f(x) - E_X[f(X)]\}\{f(y) - E_Y[f(Y)]\}dxdy \\ &= \int P_X(x)P_Y(y)\{f(x) - E_X[f(X)]\}\{f(y) - E_Y[f(Y)]\}dxdy \\ &= \int P_X(x)\{f(x) - E_X[f(X)]\}dx \int P_Y(y)\{f(y) - E_Y[f(Y)]\}dy \\ &= \{E_X[f(x)] - E_X[f(X)]\}\{E_Y[f(y)] - E_Y[f(Y)]\} \\ &= 0 \end{aligned}$$

c) Since the mean is

$$\begin{aligned} E_{X_1,\ldots,X_n}[h] &= -\frac{1}{n}\sum_i E_{X_i}[logP_X(X_i)] \\ &= -E_X[logP_X(X)] \\ &= H[X], \end{aligned}$$

the estimator is unbiased. The variance is

$$\begin{aligned} var[h] &= E_{X_1,\ldots,X_n}\{(h - E_{X_1,\ldots,X_n}[h])^2\} \\ &= E_{X_1,\ldots,X_n}\{(h - H[X])^2\} \\ &= E_{X_1,\ldots,X_n}\{(-\frac{1}{n}\sum_i \log P_X(X_i) - H[X])^2\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2} E_{X_1,\ldots,X_n}\{(\sum_i \log P_X(X_i) - H[X])^2\} \\
&= \frac{1}{n^2} E_{X_1,\ldots,X_n}\{(\sum_i \log P_X(X_i) - H[X])(\sum_j \log P_X(X_j) - H[X])\} \\
&= \frac{1}{n^2} \sum_{i,j} E_{X_i,X_j}\{(\log P_X(X_i) - H[X])(\log P_X(X_j) - H[X])\} \\
&= \frac{1}{n^2} \sum_i E_{X_i}\{(\log P_X(X_i) - H[X])^2\} \quad (\text{using } \mathbf{b})) \\
&= \frac{1}{n} E_X\{(\log P_X(X) - H[X])^2\} \\
&= \frac{1}{n} var[\log P_X(X)]
\end{aligned}
$$

Assuming that $\log P_X(X)$ has finite variance, this implies that the the variance of the estimate goes to zero assymptotically.

2. a) According to bayes decision rule, the optimal decision function, under the "0/1" loss is "pick i over j" if

$$\begin{aligned} P_{Y|X}(i|x) &> P_{Y|X}(j|x) \\ P_{X|Y}(x|i)P_Y(i) &> P_{X|Y}(x|j)P_Y(j) \\ P_{X|Y}(x|i) &> P_{X|Y}(x|j) \text{ (as the priors are equal)} \end{aligned}$$

Taking Log on both sides,

$$\begin{aligned} \log P_{X|Y}(x|i) - \log P_{X|Y}(x|j) &> 0 \\ (x-\mu_j)^T\Sigma^{-1}(x-\mu_j) - (x-\mu_i)^T\Sigma^{-1}(x-\mu_i) &> 0 \\ 2(\mu_i-\mu_j)^T\Sigma^{-1}x - (\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i-\mu_j) &> 0 \\ \Rightarrow (\mu_i-\mu_j)^T\Sigma^{-1}\left(x - \frac{1}{2}(\mu_i+\mu_j)\right) &> 0 \end{aligned}$$

So the decision boundary can be expressed as a hyperplane, $w_{ij}^T(x - x_0^{ij}) = 0$, where w_{ij} is the normal and x_0^{ij} is a point through the hyperplane, with

$$w_{ij} = \Sigma^{-1}(\mu_i - \mu_j) \quad (1)$$

$$x_0^{ij} = \frac{1}{2}(\mu_i + \mu_j) \quad (2)$$

b) Yes, we can determine w_{13} using the known quantities. Using 1, we can write

$$w_{12} = \Sigma^{-1}(\mu_1 - \mu_2) \quad (3)$$

$$w_{23} = \Sigma^{-1}(\mu_2 - \mu_3) \quad (4)$$

$$\Rightarrow w_{12} + w_{23} = \Sigma^{-1}(\mu_1 - \mu_3) \quad (5)$$

$$= w_{13} \quad (6)$$

c) No, we cannot determine the points $x_0^{12}, x_0^{23}, x_0^{13}$ using the known quantities. Using given quantities we can compute,

$$\begin{aligned} (\mu_1 - \mu_2) &= \Sigma w_{12} \\ (\mu_2 - \mu_3) &= \Sigma w_{23} \\ (\mu_1 - \mu_3) &= \Sigma w_{13} \end{aligned}$$

But as shown in part **b)** the three equations are not independent, and one relation can be obtained from the other two. So we cannot determine $x_0^{12}, x_0^{23}, x_0^{13}$, or equivalently, μ_1, μ_2, μ_3. However, we just need one of μ_1, μ_2, μ_3 to determine the other two. Geometrically, the given quantities fix the distance between the means. But an arbitrary translation to all three points will still preserve the distances. So we need atleast one point to determine the positions uniquely.

3. a) The log-likelihood is

$$l(\theta) = \sum_i \log h(x_i) + \nu(\theta) \sum_i T(x_i) - nA(\theta).$$

Setting its derivative to zero, we have

$$\frac{\partial \nu}{\partial \theta} \sum_i T(x_i) = n \frac{\partial A}{\partial \theta},$$

and the equation follows. For this solution to be a maximum we need that

$$\frac{\partial^2 l}{\partial \theta^2} = \frac{\partial^2 \nu}{\partial \theta^2} \sum_i T(x_i) - n \frac{\partial^2 A}{\partial \theta^2} < 0$$

i.e. the additional constraint that

$$\frac{\partial^2 A}{\partial \theta^2} > \frac{\partial^2 \nu}{\partial \theta^2} \frac{1}{n} \sum_i T(x_i)$$

b) According to bayes decision rule, the optimal decision function, under the "0/1" loss is "pick $i = 0$" if

$$\begin{aligned} P_{Y|X}(0|\mathcal{D}) &> P_{Y|X}(1|\mathcal{D}) \\ P_{X|Y}(\mathcal{D}|0)P_Y(0) &> P_{X|Y}(\mathcal{D}|1)P_Y(1) \\ P_{X|Y}(x_1, \ldots, x_n|0)P_Y(0) &> P_{X|Y}(x_1, \ldots, x_n|1)P_Y(1) \end{aligned}$$

Taking the log on both sides, and using the independence assumption,

$$\begin{aligned} \log \frac{P_{X|Y}(x_1, \ldots, x_n|0)}{P_{X|Y}(x_1, \ldots, x_n|1)} + \log \frac{P_Y(0)}{P_Y(1)} &> 0 \\ [\nu(\theta_0) - \nu(\theta_1)] \sum_i T(x_i) - n[A(\theta_0) - A(\theta_1)] + \log \frac{\pi_0}{\pi_1} &> 0 \\ s_n &> \frac{1}{[\nu(\theta_0) - \nu(\theta_1)]} \left([A(\theta_0) - A(\theta_1)] + \frac{1}{n} \log \frac{\pi_1}{\pi_0} \right), \end{aligned}$$

where we have assumed that $\nu(\theta_0) \geq \nu(\theta_1)$. Therefore, the threshold is, $T = \frac{1}{[\nu(\theta_0) - \nu(\theta_1)]} \left([A(\theta_0) - A(\theta_1)] + \frac{1}{n} \log \frac{\pi_1}{\pi_0} \right)$.

c)

$$\begin{aligned} KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)] &= E_{X|Y} \left[\log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} \middle| Y = 1 \right] \\ &= [\nu(\theta_1) - \nu(\theta_0)] E_{X|Y} [T(x)|Y = 1] - [A(\theta_1) - A(\theta_0)] \\ &= -\left\{ [\nu(\theta_0) - \nu(\theta_1)] E_{X|Y} [T(x)|Y = 1] - [A(\theta_0) - A(\theta_1)] \right\} \\ KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)] &= E_{X|Y} \left[\log \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} \middle| Y = 0 \right] \\ &= [\nu(\theta_0) - \nu(\theta_1)] E_{X|Y} [T(x)|Y = 0] - [A(\theta_0) - A(\theta_1)] \end{aligned}$$

d) Note that the Bayes decision rule is to pick class $i = 0$

$$[\nu(\theta_0) - \nu(\theta_1)] \frac{1}{n} \sum_i T(x_i) - [A(\theta_0) - A(\theta_1)] > \frac{1}{n} \log \frac{\pi_1}{\pi_0}$$

and class 1 otherwise. As n goes to infinity this becomes

$$[\nu(\theta_0) - \nu(\theta_1)] E_X [T(x)] - [A(\theta_0) - A(\theta_1)] > 0$$

where X is the random variable from which the sample $\mathcal{D}$ to classify was drawn. If the sample comes from class 0, then $E_X [T(x)] = E_{X|Y} [T(x)|Y = 0]$. If it came from class 1, then $E_X [T(x)] = E_{X|Y} [T(x)|Y = 1]$. Hence, the left-hand side of the BDR becomes

$$KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)]$$

for a sample from class 0 and

$$-KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)]$$

for a sample from class 1. Since the KL divergence is never negative, this implies that class 0 will always be chosen when the sample comes from class 0 and class 1 will always be chosen for when the sample comes from class 0. This shows that the classifier has zero error.

e) For class 0 everything stays the same, and the decision is always correct. For class 1, $\frac{1}{n} \sum_i T(x_i)$ converges to

$$\epsilon E_{X|Y} [T(x)|Y = 0] + (1 - \epsilon) E_{X|Y} [T(x)|Y = 0],$$

and the left hand side of the BDR becomes

$$\epsilon KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)] - (1 - \epsilon) KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)].$$

Hence, there is no error when

$$\frac{\epsilon}{1 - \epsilon} \leq \frac{KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)]}{KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)]}.$$