10-704 Information Processing & Learning

Quiz 2 Solutions *April 28^{th} 2015*

Short Questions

(5 Points $\times$ 8 = 40 Points)

1. We wish to encode a dictionary of 5 symbols $\{a, b, c, d, e\}$ using a ternary alphabet $\{0, 1, 2\}$. Identify the following 5 codes as **S**: Singular, **NS**: Nonsingular but not uniqely decodable, **UD**: Uniquely decodable but not instantaneous, and **I**: Instantaneous

   (a) $\{0, 1, 2, 0, 1\}$

   (b) $\{01, 10, 11, 02, 2\}$

   (c) $\{0, 1, 11, 21, 02\}$

   (d) $\{0, 21, 02, 2, 21\}$

   (e) $\{000, 1112, 1111, 2222, 2221\}$

2. Let $Y = X_1 + X_2$ where X_1, X_2 are not necessarily indepdendent and satisfy $\mathbb{E}X_i^2 \leq P$ for $i = 1, 2$. Find the maximum entropy of Y.

3. State True/ False.
   (a) The Jeffrey's prior is invariant to reparametrization.
   (b) Reference priors are invariant to reparametrization in one dimension but not in more than one dimension.
   (c) The redundancy-capacity theorem tells us that the reference prior is the worst-case prior achieving minimax risk in learning a parameter θ from data X.
4. The exponential family of distributions parametrized by θ is characterized via the pdf

$$p_\theta(x) = h(x) \exp\left(\sum_{k=1}^{s} \eta_k(\theta) T_k(x) - B(\theta)\right)$$

   You have n samples $\{X_1, \ldots, X_n\}$, from the above distribution. Indicate whether the following statistics are sufficient ? (You may circle the sufficient statistics.)
   (a) $\{X_1, \ldots, X_n\}$
   (b) $\{\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_s(X_i)\}$
   (c) $\{\sum_{i=1}^n \sum_{k=1}^s T_k(X_i)\}$
   (d) $\{\prod_{i=1}^n h(X_i), \sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_s(X_i)\}$
   (e) $\{\prod_{i=1}^n h(X_i), \sum_{i=1}^n \sum_{k=1}^s T_k(X_i)\}$
5. Consider the density given below. Note that this is the $\Gamma(2, \theta)$ distribution.

$$p_\theta(x) = \frac{1}{2\theta^2} x \exp\left(\frac{-x}{\theta}\right) \mathbb{1}(x > 0)$$

   What is the Cramer-Rao lower bound on the variance of any unbiased estimator for θ.
   Hint: The mean of a $\Gamma(\alpha, \beta)$ distribution is $\alpha\beta$.

6. Consider the distribution $p = \{1/2, 1/4, 1/8, 1/8\}$ on symbols $\{a, b, c, d\}$. What is the Shannon-Fano-Elias Code for the sequence acb when each symbol is drawn i.i.d from p ?

7. Let $\mathcal{V} = \{-1, 1\}^d$, and let $\theta(v) = v$. Which of the following losses satisfy the decomposability requirement for Assouad's Method? You may circle your answers.

   (a) Squared loss: $l(\theta, \theta') = \|\theta - \theta'\|_2^2$.

   (b) ℓ_1 loss: $l(\theta, \theta') = \|\theta - \theta'\|_1$.

   (c) ℓ_∞ loss $l(\theta, \theta') = \max_{j \in [d]} |\theta_j - \theta'_j|$.

8. Given n i.i.d. samples $X_i \in \{+1, -1\}$ from a distribution $P \sim \text{Bernoulli}(1/2)$, Sanov's theorem states that $P(\sum_{i=1}^n X_i > n/2)$ decays asymptotically as which of the following:

   (a) $2^{-nD((3/4,1/4)\|(1/2,1/2))}$

   (b) $2^{-nD((3/4,1/4)\|(1/4,3/4))}$

   (c) $2^{-nD((1/2,1/2)\|(3/4,1/4))}$

Solutions

1. (a) S

   (b) I

   (c) NS

   (d) S

   (e) I

2. Noting that $\mathbb{V}(Y) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\mathbb{C}\mathrm{ov}(X_1, X_2) \leq \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\sqrt{\mathbb{V}(X_1)\mathbb{V}(X_2)} \leq 4P$. Therefore $H(Y) \leq \frac{1}{2}\log(8\pi eP)$. This upper bound is achievable when $X_1 = X_2$ and is sampled from $\mathcal{N}(0, P)$.

3. (a) T

   (b) F

   (c) T

4. (a), (b) and (d)

5. $\theta^2/2$

6. 0110101

7. (a) Yes

   (b) Yes

   (c) No

8. Ans: (a)

Long Questions

1. **(5+10+5 Points) Rate Distortion and Identifying Anomalies**

   In this problem, you will pose the problem of identifying anomalous points as a rate-distortion problem. Consider data X that we would like to map to T such that T is w if data X is "non-anomalous" (similar to other data points) and X if X is "anomalous" (different from other data points). Here, w is a fixed value indicating that the data was non-anomalous. You may assume that the distribution of X is known.

   (a) Pose it as a rate-distortion problem where the distortion is $\|X - T\|^2$.

   (b) Write down the iterative steps in Blahut-Arimoto algorithm for finding the rate-distortion function starting from a guess of initial probabilities $p^{(0)}(T = w)$ and $p^{(0)}(T = x)$. You don't need to derive it from scratch. At iteration $i = 1, 2, \ldots,$

   $$p^{(i)}(T = w|X = x) =$$

   $$p^{(i)}(T = x|X = x) =$$

   Then update

   $$p^{(i)}(T = w) =$$

   $$p^{(i)}(T = x) =$$

   (c) Show that the optimal value of w corresponds to the expectation of X conditioned on it being mapped to non-anomalous.

Solution

(a) $\min_{p(t|x)} I(X,T) \;\; s.t. \;\; \mathbb{E}[||X-T||^2] \leq D$

(b) $p^{(i)}(T=w|X=x) \propto p^{(i-1)}(T=w)e^{-\beta\|x-w\|^2}$
$p^{(i)}(T=x|X=x) \propto p^{(i-1)}(T=x)$

Then update
$p^{(i)}(T=w) = \sum_x p^{(i)}(T=w|X=x)p(x)$
$p^{(i)}(T=x) = p^{(i)}(T=x|X=x)p(x)$

(c) $\frac{\partial}{\partial w}\mathbb{E}[\|X-T\|^2] = \frac{\partial}{\partial w}\sum_x \|x-w\|^2 p(t=w|x)p(x) = -\sum_x 2(x-w)p(t=w|x)p(x) = 0$. This implies

$$w = \frac{\sum_x xp(t=w|x)p(x)}{\sum_x p(t=w|x)p(x)} = \sum_x xp(x|t=w)$$

i.e. it corresponds to the expectation of X conditioned on it being mapped to non-anomalous.

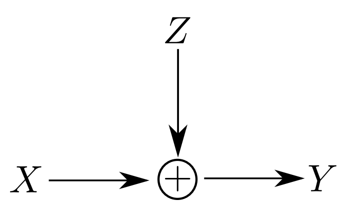
Here is an alternative solution,

$$\mathbb{E}\|X-T\|^2 = \int_A (x-w)^2 p(x)$$
$$\frac{\partial}{\partial w}\mathbb{E}\|X-T\|^2 = \frac{\partial}{\partial w}\int_A \|X-w\|^2 p(x) \implies w = \frac{\int_A xp(x)}{p(A)} = \mathbb{E}[X|X \in A]$$

Here A is the non-anomalous region. The calculation assumes that A is known.

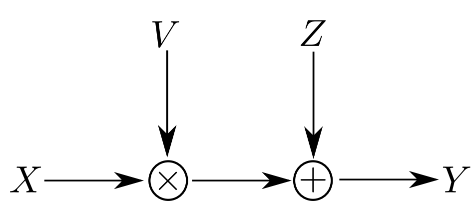
2. **(10+10 Points) Channel Capacity**

(a) Consider the additive channel below, where $X \in \mathcal{X} = \{-2,-1,0,1,2\}$ and the output is $Y = X + Z$. Z is noise uniformly distributed over $[-1,1]$ and is independent of X.



Calculate the capacity $C = \max_{p(x)} I(X;Y)$ of this channel and describe the distribution $p(x)$ used to achieve that capacity.

(b) Now consider the following channel where the output is $Y = VX + Z$ where V, Z are random variables independent of X. All X, Y, V and Z are scalars.



Let the capacity of the channel when V is known be $C_V = \max_{p(x)} I(X;Y|V)$ and when V is unknown be $C = \max_{p(x)} I(X;Y)$. Prove that $C_V \geq C$.

Solution

(a) Write $I(X;Y) = H(Y) - H(Y|X) = H(Y) - 1$ bits. Since Y is a distribution over $[-3,3]$ its entropy is at most $\log 6$ bits achieved by a uniform distribution. This can be achieved via the prior $\{1/3, 0, 1/3, 0, 1/3\}$. Hence, $C = \log 3$ bits.

(b) For this, we write the mutual information $I(X;V,Y)$ in two ways via the chain rule,

$$I(X;V,Y) = I(X;V) + I(X;Y|V) = I(X;Y) + I(X;V|Y)$$

As $I(X;V) = 0$ due to independence we have $I(X;Y|V) \geq I(X;V)$. The statement follows.

3. **(20 pts)** Consider the following simple model for similarity based clustering. There are n objects and they are partitioned into two sets of size $n/2$. Call one of the sets S, so that the other is S^C.

You observe an $n \times n$ matrix M with $M_{ij} \sim \mathcal{N}(\gamma, 1)$ if $i,j \in S$ or $i,j \in S^C$ and with $M_{ij} \sim \mathcal{N}(-\gamma, 1)$ otherwise. Given this matrix, you would like to recover the set S and the set S^C.

An estimator T outputs two sets (A,B) and we say that $(A,B) = (A',B')$ if either $A = A'$ and $B = B'$ or $A = B'$ and $B = A'$. This just means that the clustering found by T agrees with the true clustering.

Show that the minimax risk:

$$\inf_T \sup_{S \subset \{1,\ldots,n\}, |S| = n/2} \mathbb{P}_S[T(M) \neq (S, S^C)],$$

is lower bounded by a constant when $\gamma \leq c\sqrt{\frac{\log(n)}{n}}$. You need not explicitly track the constant factors in your calculations.

Hint: Use Fano's Inequality. Fix one partition (S, S^C) of $n/2$ objects in each cluster and an element $i \in S$. Consider a discretization of the hypothesis space that includes this clustering along with all $n/2$ clusterings based on swapping element i with an element from S^C.

Solution: Fix (S_0, S_0^C) to be a clustering where each set has $n/2$ elements. Fix one element i in S_0 and for each element $j \in S_0^C$ let S_j be the clustering that swaps i and j. Clearly there are $n/2$ such alternatives, and each one disagrees with (S_0, S_0^C) on $\Theta(n)$ similarities.

The KL is $\Theta(n\gamma^2)$ and the entropy is $\Theta(\log(n))$ so by Fano's inequality:

$$P_e \geq 1 - c\frac{n\gamma^2 + \log 2}{\log n}$$

which is bounded away from zero when $\gamma \leq \sqrt{\frac{\log n}{n}}$.