10-704 Information Processing & Learning

Quiz 2 *April 28th 2015*

Name: : _____

Andrew Id: : _____

Department: : _____

Guidelines

1. **Please don't turn this page until instructed.**
2. Write your name, Andrew Id and department in the space provided above.
3. You have seventy-five **(75)** minutes for this exam.
4. This exam has **seven (7)** pages on seven sheets of paper including this page.
5. This exam has 2 Sections. The first section has 8 short questions and the second section has 3 long questions. The number of points allocated for each question is indicated next to the question. The total number of points is 100.
6. This exam is **open book**. You may use any material such as cheat sheets, class notes etc. No electronic devices are permitted.
7. The questions vary in difficulty. The points allocated per question **do not** entirely reflect the level of difficulty. Do not spend too much time on one question.
8. If any question is unclear, you may write your own interpretation and answer the question.
9. The questions appear only on one side of the paper. You may use the other side for rough work. If you still need extra paper, please ask one of the instructors.

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

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

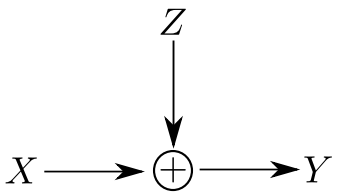
   $$p^{(i)}(T = w) =$$

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

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

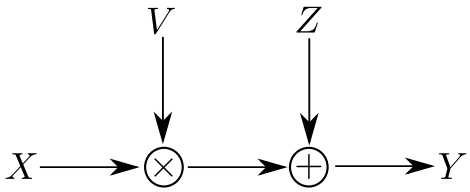
2. **(10+10 Points) Channel Capacity**

(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 $C_V \geq C$.

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.