Statistical Information Theory MT7037 VT20
Take home project 2

Examiner: Chun-Biu Li (cbli@math.su.se)
Posting date: Mar 12, 2020 Deadline: **8pm Mar 22, 2020**

***** Read carefully the following instructions before starting *****

- The submission should include your report to the questions (a single pdf file) and the signed confirmation sheet. The source code in Exercise 2c should be attached as appendices in your report. The files should be submitted at the course page.
- Total points is 60 in this project. A 9-point deduction will be imposed for late submission.
- Only the reports of LADOK registered students, except for PhD students, will be graded.
- Credit points will be given based on clear logical explanation and steps leading to the final solution. Also redundant writing irrelevant to the solution can result in point deduction.
- The report, not including the appendices, should be ≤ 8 pages.
- The report must be completed independently. Plagiarism or other forms of cheating is a serious act. To underline this, the *signed confirmation sheet* must be submitted to declare that your work is made in accordance with the rules for written exams at Stockholm University (see course page).

Exercise 1 (Total: 21p)

To better understand the concepts of maximum entropy distribution and multi-information, this exercise requires you to work out and generalize some calculations in the two papers by Schneidman *et al.*, (*Phys. Rev. Lett.*, 91:238701 2003 and *Nature*, 440:1007 2006).

a) For N binary variables $x_1, x_2, \cdots, x_N$ taking values ± 1 (instead of 0 or 1 as discussed in the class) with only the pairwise marginals known, derive the expression for the maximum entropy distribution. Does your answer still look like Eq. 1, namely, the Ising model, in Schneidman *et al.*, *Nature*, 440:1007 2006? (7p)

b) Derive all the numerical values shown in Fig. 1 in Schneidman *et al.*, *Phys. Rev. Lett.*, 91:238701 2003 for the case of XOR function. (6p)

c) As discussed in class, the maximum entropy distribution $\tilde{P}^{(k)}(x_1, x_2, \cdots, x_N)$ for N binary variables (take values 0 or 1) constrained by the kth ($k \leq N$) order marginals has the form:

$$\exp\left[\theta_0 + \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \cdots + \sum_{i<j<\cdots<l} \theta_{ij\cdots l} x_i x_j \cdots x_l\right],$$

where $x_i x_j \cdots x_l$ is a product of k factors. Now consider the XOR case as in part b) with $N = 3$ variables where all the Nth marginals are known, i.e., all numerical values of the full joint probability $P(x_1, x_2, x_3)$ are known. Can you solve for θ_i, θ_{ij} and θ_{ijk} in this case? If yes, show their numerical values; If not, explain what is the problem and how to resolve it. (8p)

Exercise 2 (Total: 39p)

a) Consider the minimization of the Lagrange function $\mathcal{F} = I(X, \tilde{X}) + \beta \langle d(x, \tilde{x}) \rangle_{p(x,\tilde{x})}$ with respect to $p(\tilde{x}|x)$ in the rate distortion theory, show with detailed steps that the formal solution is given by:

$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp[-\beta d(x, \tilde{x})]$, with $Z(x, \beta) = \sum_{\tilde{x}'} p(\tilde{x}') \exp[-\beta d(x, \tilde{x}')]$. (6p)

Parts b) to e) of this exercise require you to develop the code for the rate distortion theory and examine its performance in non-parametric information-based clustering problem. To begin with, you need to first download the date file (Data-Exercise2-2020) from the moodle course page that contains 200 data points to be clustered. In the file, the first and second columns are the x- and y-coordinates of the data points, respectively. Moreover, the first and second 100 points are independently sampled from the normal distributions, $N\left(\mu = (0 \quad 1), \Lambda = \begin{pmatrix} 0.09 & 0 \\ 0 & 0.09 \end{pmatrix}\right)$ and $N\left(\mu = (1 \quad 0), \Lambda = \begin{pmatrix} 0.09 & 0 \\ 0 & 0.09 \end{pmatrix}\right)$, respectively.

b) Plot the data points on the x-y plane and prepare the element-to-element distance matrix $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ $(i, j = 1, 2, \cdots, 200)$ that will be the input for the clustering algorithm. Also assume each data point carries the same weight, i.e., $p(\mathbf{x}_i) = 1/200$ for all i. (2p)

c) Write a code (no restriction on the program language) to implement the Blahut-Arimoto algorithm to evaluate the clustering membership probability, $p(\tilde{x}|\mathbf{x})$, with fixed number of clusters, N_c, and compression-distortion tradeoff parameter, β. Your code should implement a multiple run each starting with random initial conditions. **Note:** Your source code should include clear comments/documentations to describe what are evaluating. I may later randomly ask a few students, especially those without clear documentations, to demonstrate how their code works. (18p)

d) Run your code to construct the information curve for $N_c = 2, 3, 4$. Hint: Choose different values of β in between 1 to 40. (7p)

e) As we have already known that the correct number of clusters is 2, propose a reasonable way using the quantities evaluated from your code (e.g. $I(\mathbf{X}, \tilde{X})$,$\langle d(\mathbf{x}, \tilde{x}) \rangle_{p(\mathbf{x},\tilde{x})}$, the Lagrange function, etc.) and the information curves to correctly identify the number of clusters. You should clearly explain your rationale and state explicitly which quantities, graphs and/or curves are used in the identification. (6p)

Good Luck!