Advanced Machine Learning (GR5242)
Fall 2017

Homework 2

Due: Thurday 5 October, at 4pm (for both sections of the class)

Homework submission: Please type or **neatly** write up your solutions to the following exercises and submit a pdf of the results to courseworks. Illegible solutions will be assumed to be incorrect, so take the time to clean up your work before scanning anything handwritten.

Note: The problems marked with a $*$ (problems 2 and 3) are former exam problems; they are related to the material in class, but also meant to give you an idea of what midterm questions may look like.

Problem 1 (Variational inference for the Potts model)

We consider random variables $X_1, \ldots, X_n$ with value in $\{-1, 1\}$, with joint mass function

$$P(x_1, \ldots, x_n) = \frac{1}{Z(\beta)} e^{-\beta G(x_1, \ldots, x_n)}$$

(The function G was denoted H in class; we use G here to avoid confusion with the entropy $\mathbb{H}$.) G is given by

$$G(x_1, \ldots, x_n) = -\sum_{i,j=1}^{n} w_{ij} x_i x_j - \sum_{i=1}^{n} h_i x_i \,, \tag{1}$$

where w_{ij} and h_i are constants. Our goal is to solve the variational approximation problem

$$Q^* = \arg\min_{\tilde{Q} \in \mathcal{Q}} D_{\mathsf{KL}}(\tilde{Q} \| P) \,.$$

The class $\mathcal{Q}$ of approximating distributions consists of factorial distributions

$$Q\left(x_1, ..., x_n\right) = \prod_{i=1}^{n} Q_i\left(x_i\right) \qquad \text{where} \qquad Q_i\left(x_i\right) = \begin{cases} q_i & x_i = 1 \\ (1 - q_i) & x_i = -1 \end{cases} \text{ for some } q_i \in [0, 1] \,, \tag{2}$$

so each factor Q_i is a Bernoulli distribution on $\{-1, +1\}$. It will turn out convenient to rewrite the factors Q_i as

$$Q_i\left(x_i\right) = \begin{cases} \frac{1+m_i}{2} & x_i = 1 \\ \frac{1-m_i}{2} & x_i = -1 \,, \end{cases}$$

with $m_i \in [-1, +1]$. The marginal Q_i is completely specified by the parameter m_i, and so is Q by $\mathbf{m} = (m_1, ..., m_n)$. The class $\mathcal{Q}$ of approximating distributions is therefore

$$\mathcal{Q} = \left\{ Q(\mathbf{m}) = \prod_{i=1}^{n} Q_i(x_i) \;\middle|\; m_i \in [-1, +1] \right\} \,.$$

a) Show that the Kullback-Leibler divergence is given by

$$D_{\mathsf{KL}}(Q \| P) = \mathbb{H}(Q) + \beta \mathsf{E}_Q \left[-\beta G(x_1, \ldots, x_n) \right]$$

where $\mathbb{H}$ is the entropy and $\mathsf{E}_Q[\bullet]$ denotes expectation with respect to Q.

Since β is fixed and strictly positive, minimizing $D_{\mathsf{KL}}(Q\|P)$ is equivalent to minimizing $\frac{1}{\beta}D_{\mathsf{KL}}(Q\|P)$, and it is customary to define the variational cost function as

$$F(Q) \quad := \quad \frac{1}{\beta}D_{\mathsf{KL}}(Q\|P) \quad = \quad \frac{1}{\beta}\mathbb{H}(Q) + \mathsf{E}_Q\left[G\right]$$

The distribution which minimizes F is computed by solving the optimization problem

$$\mathbf{m}^* = \arg\min_{\mathbf{m}} F\left(Q(\mathbf{m})\right) \ . \tag{3}$$

We start by computing some expected values that will be needed to set up the function F. (Before you start, recall what the joint expectation behaves like for independent random variables.)

b) Can you show that

$$\sum_{\mathbf{x}\in\{-1,1\}^n} Q\left(\mathbf{x}\right) x_i \ = \sum_{x_i\in\{-1,+1\}} Q_i\left(x_i\right) x_i \tag{4}$$

holds for any factorial distribution Q? From (4), you can immediately deduce $\mathsf{E}_Q\left[x_i\right] = m_i$.

c) Show that the expected value of G in (1) is

$$\mathsf{E}_Q\left[G\right] = -\sum_{i,k=1}^{n} w_{ik}m_i m_k - \sum_{i=1}^{n} m_i h_i \ . \tag{5}$$

d) Show that the entropy of Q is

$$\mathbb{H}[Q] = -\sum_i \left(\frac{1+m_i}{2}\log\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2}\log\left(\frac{1-m_i}{2}\right)\right) \ . \tag{6}$$

These are the steps which make explicit use of the factorial assumption (2). Take another look at your computations: In each of them, you will find at least one step that would have been considerably more complicated if we had not assumed a factorial distribution. (This is for your interest and will not be graded).

With (5) and (6), F takes the form

$$\begin{aligned} F\left(Q(\mathbf{m})\right) \ = \ & -\sum_{i,k=1}^{n} w_{ik}m_i m_k - \sum_{i=1}^{n} m_i h_i \\ & + \frac{1}{\beta}\sum_i \left(\frac{1+m_i}{2}\log\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2}\log\left(\frac{1-m_i}{2}\right)\right) \ . \end{aligned}$$

What remains to be done is to find the probability distribution which optimizes the function F by solving (3):

e) Show that the optimal distribution Q^* is given by parameter values satisfying

$$m_i = \tanh\Big(\beta\Big(\sum_{j=1}^{n} w_{ij}m_j + h_i\Big)\Big) \ . \tag{7}$$

Hint: Take derivatives for one m_i at a time, keeping the others constant. Recall $\operatorname{arctanh}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$.
For given parameter values w_{ij} and h_i, (7) then defines a set of nonlinear equations that has to be solved to obtain a specific solution Q^* (which you are not asked to do here).

Problem 2* (HMMs and topics)

Suppose we have to model text data which is streamed from a news feed; each news item is part of a single topic. After a (random) number of words, the new item ends and the next item begins, which in general has a different topic. Over time, topics may repeat. Suppose we have estimated empirically that:

- There are K topics, and we have estimated the probability vectors $\theta_1, \ldots, \theta_K$ (where θ_k is the parameter vector of a multinomial which models text with topic k).
- At any given word, the probability of remaining within the current topic is 0.99.
- The probability of switching to a different topic is 0.01. For simplicity, assume all topics are equally probable, so the probability to switch to a specific new topic is $q := \frac{0.01}{K-1}$ for each topic.

a) Define a hidden Markov model to model the word sequence $X_1, X_2, \ldots$. Please make sure that you specify:

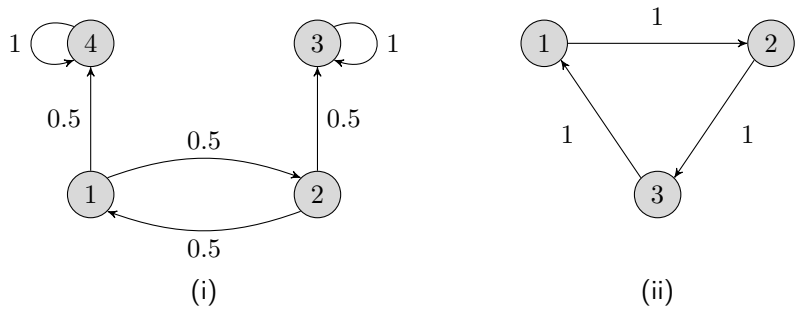
- The state space of your model.
- The observed and hidden variables.
- The transition and emission probabilities.

b) Are the variables X_i and X_{i+2} (for any $i \in \mathbb{N}$) stochastically dependent?

Problem 3* (Markov Chains)

This problem revisits the conditions in the fundamental theorem on Markov chains, in the (simpler) case of finite state spaces. Even the finite state space case is relevant to Markov chain sampling—if we sample from a Potts model with n vertices, for example (although the size of the state space is then 2^n, rather than 3 or 4 as below).

a) Does either of the following two Markov chains have invariant distributions? If so, is it unique? Please explain your answers.



b) The *gambler's fallacy* is the belief among gamblers that, when playing a game of chance for money, an "unlucky streak" will be followed by luck. Suppose that were true and a gambler could be lucky, unlucky or neutral. If the outcomes of n successive games are variables $X_1, \ldots, X_n$, does the gambler's fallacy constitute an i.i.d. model, a Markov model, or a hidden Markov model for the sequence $X_1, \ldots, X_n$? Please explain your answer.