

STAT 538 Homework 4  
Out February 20, 2008  
Due February 28, 2008  
©Marina Meilă  
mmp@stat.washington.edu

Reading: B&V 3.1-3.3

**Problem 1 – The conjugate of the entropy**

In this problem you will compute the conjugate of the entropy of a discrete distribution.

**1.** Assume that the domain of  $X$  is  $\mathcal{X} = \{0, 1\}$ , so that any distribution on  $\mathcal{X}$  is parametrized by  $p = Pr[X = 1]$ . The entropy is a concave function, so we will compute the conjugate of the negative entropy  $f(p) = -H(p)$  with  $H(p) = -p \ln p - (1 - p) \ln(1 - p)$ .

**1.1** Compute the gradient of  $f$

**1.2** Solve the equation  $f'(p) = q$ . What is the domain of  $q$ ?

**1.3** Now compute the expression of the dual  $f^*(q)$  as a function of  $q$  and verify that it is convex.

**2.** Now let  $\mathcal{X} = \{1, \dots, m\}$  and let  $p = (p_1, \dots, p_m)$  parametrize a distribution on it. Evaluate the dual of  $f(p) = -H(p)$  in this case for  $H(p) = -\sum_{i=1}^m p_i \ln p_i$ .

Note that one of the variables  $p_i$  is redundant being completely determined by the other  $m - 1$ .

**Problem 2 – A variational bound for the logistic function**

Just like the term “spectral” means “having to do with eigenvalues and eigenvectors, or an eigendecomposition”, the term *variational* means “having to do with an optimization problem”. In this course, the “optimization problem” will typically be the “sup” in the definition of the convex conjugate.

More to the point, if a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is convex, then we have

$$f(x) \geq yx - f^*(y) \tag{1}$$

which, for any  $y$ , represents a linear lower bound for  $f$ . Such a bound is called

a variational bound. We will try to play a similar trick for the logistic function, which is neither convex, nor concave.

1. Denote the logistic function by  $g$

$$g(x) = \frac{e^x}{1 + e^x}$$

Show that  $g$  is log-concave.

2. Let  $f = -\ln g$ . Obtain the expression of  $f^*$ , the conjugate of  $f$ , and its domain.

3. Use equation (1) to obtain a lower bound on  $f$ . From the obtained bound on  $f$ , get a bound on the logistic function  $g$ . Is this a lower or an upper bound?

4. **Plots:** Display on a graph the function  $f$  and the linear bounds for  $y = -0.2, -0.5, -0.7, -0.95$ . On a separate graph display  $g$  and the variational bounds corresponding to  $y = -0.2, -0.5, -0.7, -0.95$ .

### Problem 3 – Divergence minimization

a. Let  $V = \{X_1, X_2, \dots, X_m\}$  be a set of binary variables and  $P, Q$  be probability distributions over this domain.  $Q$  is a *completely factored* distribution, i.e

$$Q(X_1, X_2, \dots, X_m) = \prod_{j=1}^m Q_j(X_j)$$

We assume  $P$  fixed and try to optimize  $Q$  as to best fit  $P$  by

$$Q = \operatorname{argmin}_{Q \text{ factored}} KL(P||Q)$$

where  $KL$  denoted the KL divergence.

$$KL(P||Q) = \sum_{X_1, X_2, \dots, X_m} P \ln \frac{P}{Q}$$

Prove that the optimal  $Q$  is the product of the marginals of  $P$ , in other words

$$Q_j = \sum_{X_i: i \neq j} P$$

b. Now let us examine the triangular graphical model from Minka: The variables  $X, Y, Z \in \{0, 1\}$  and  $p = p_X p_Y p_Z p_{XY} p_{XZ} p_{YZ}$  where each factor is a (possibly unnormalized) table with non-negative entries.

We approximate the factors of  $p$  by a factored  $q$  as follows:

$$q = p_X p_Y p_Z \underbrace{q_X^{XY} q_Y^{XY}}_{q_{XY}} \underbrace{q_X^{XZ} q_Z^{XZ}}_{q_{XZ}} \underbrace{q_Y^{YZ} q_Z^{YZ}}_{q_{YZ}}$$

i.e the single variable factors in  $p$  are not approximated, and the two variable factors are approximated by a product of two single variable factors.

You will write explicitly the *local divergence minimization* of Minka for this setup and for the case of the KL divergence. Denote by  $q^{\setminus XY}$  the product of all factors in  $q$  except for  $q_{XY}$ .

Write the expression of the local divergence

$$KL(p_{XY} q^{\setminus XY} || q_{XY} q^{\setminus XY}) \tag{2}$$

and bring it to a simple form.

The expression (2) is the update equation for factor  $q_{XY}$ . After having simplified it, find  $q_X^{XY}, q_Y^{XY}$  that minimize (2) when all the other factors in  $q$  and  $p$  are held constant.

**c. [Optional-extra credit]** The updating of  $q_{XY}$  is described in Minka as “message-passing”. Look at the expression (2) again from this angle. Since  $q_{XY}$ , the approximant of  $p_{XY}$  is being updated, then the information used should come from: the  $XY$  “factor node” itself, and from it’s neighbors, the “variable nodes”  $X, Y$ . (In the triangular graph in Minka, the factor node  $XY$  is the small square, having as neighbors the circles  $X, Y$ .)

Can you identify in any of the equivalent forms of the expression (2) that you have derived which are the “messages” from  $X$  and  $Y$ ?