4th homework set, Due May 8, note that the deadline of the 5th homework will be May 15

(The sum of the points is 11, i.e., if you provide a good solution for all the exercises, you get one extra point)

1. (1p.) For a closed convex set Π of distributions on A, show that P^* maximizes $H(\mathbb{P})$ subject to $\mathbb{P} \in \Pi$ if and only if P^* is the I-projection of the uniform distribution on A onto Π, and that then

$$D(\mathbb{P}||\mathbb{P}^*) \leqslant H(\mathbb{P}^*) - H(\mathbb{P}), \text{ for all } \mathbb{P} \in \Pi. \quad (1)$$

2. (4p.) Let Ξ be the log-linear family of distributions on $\Omega = \bigtimes_{i=1}^{d} \{1, \ldots, r_i\}$ with interactions $\gamma \in \Gamma$ where $\Gamma = \{\{1,2\},\{2,3\},\ldots,\{d-1,d\}\}$ (taking for Q the uniform distribution on Ω).

   (a) Show that $\mathbb{P} \in \mathcal{P}(\Omega)$ with $S(\mathbb{P}) = \Omega$ belongs to Ξ if and only if it corresponds to a Markov chain, i.e., it equals the joint distribution of random variables X_1,..., X_d such that for each $3 \leqslant j \leqslant d$ the conditional distribution of X_j on the condition $X_1 = x_1$,..., $X_{j-1} = x_{j-1}$ does not depend on x_1, ..., x_{j-2}.
   Hint: Show first the following two statements:

   - If $\mathbb{P} \in \Xi$, i.e., $\text{Prob}(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d-1} B_i(x_i, x_{i+1})$, then for X_1,..., X_d with joint distribution $\mathbb{P}$

   $$\text{Prob}(X_d = x_d | X_1 = x_1, \ldots, X_{d-1} = x_{d-1}) = \frac{\text{Prob}(X_1 = x_1, \ldots, X_d = x_d)}{\sum_{x'_d \in \{1,\ldots,r_d\}} \text{Prob}(X_1 = x_1, \ldots, X_d = x'_d)}$$

   does not depend on x_1, ..., x_{d-2}.
   - The $\{1, \ldots, d-1\}$ marginal of P, given by the sum in the denominator above, belongs to the log-linear family of distributions on $\Omega' = \bigtimes_{i=1}^{d-1} \{1, \ldots, r_i\}$ with interactions $\{1,2\}$, ..., $\{d-2, d-1\}$.

   (b) Draw the conclusion that among all distributions $\mathbb{P} \in \mathcal{P}(\Omega)$ with prescribed marginals $\mathbb{P}^{1,2}$, $\mathbb{P}^{2,3}$, ..., $\mathbb{P}^{d-1,d}$, that with largest entropy H($\mathbb{P}$) is the joint distribution of the Markov chain X_1,..., X_d with X_i, X_{i+1} having joint distribution $\mathbb{P}^{i,i+1}$, for $i = 1, \ldots, d-1$.
   Hint: Use Problem 1

3. (2p.)

Relative entropy is cost of miscoding. Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions $p(x)$ and $q(x)$ on this random variable.

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	$\frac{1}{2}$	$\frac{1}{2}$	0	0
2	$\frac{1}{4}$	$\frac{1}{8}$	10	100
3	$\frac{1}{8}$	$\frac{1}{8}$	110	101
4	$\frac{1}{16}$	$\frac{1}{8}$	1110	110
5	$\frac{1}{16}$	$\frac{1}{8}$	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$, and $D(q||p)$.

(b) The last two columns represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus, C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy p?

(d) What is the loss if we use code C_1 when the distribution is q?

4. (4p.) (Glimpse into the general theory)

Read the general definition of the KL divergence below!

More generally, if P and Q are probability measures over a set $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the Kullback–Leibler divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP,$$

If P is not absolutely continous with respect to Q, then the KL divergence is defined to be infinity

General definition of the KL divergence from wikipedia

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of P with respect to Q, and provided the expression on the right-hand side exists. Equivalently (by the chain rule), this can be written as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) \frac{dP}{dQ}\, dQ,$$

which is the entropy of Q relative to P. Continuing in this case, if μ is any measure on $\mathcal{X}$ for which $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ exist (meaning that p and q are absolutely continuous with respect to μ), then the Kullback–Leibler divergence from Q to P is given as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

The general KL divergence is also always nonnegative, and equals 0 iff the measures P and Q are equal

Let $\mathcal{E}$ be an exponential family of distributions P_θ, $\theta = (\theta_1, \ldots, \theta_k) \in \mathcal{H}$ on an arbitrary measurable space $(\mathcal{X}, \mathcal{F})$, defined by

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x) = \mathrm{e}^{-\Lambda(\theta)+\sum_{j=1}^k \theta_j f_j(x)}, \quad \Lambda(\theta) = \ln \int \mathrm{e}^{\sum_{j=1}^k \theta_j f_j(x)} \mu(\mathrm{d}x),$$

where $f_1, \ldots, f_k$ are given (measurable) functions on $(\mathcal{X}, \mathcal{F})$ and $\mathcal{H} = \{\theta : \Lambda(\theta) < \infty\}$.

Given a sample $x = (x_1, \ldots, x_n)$ the MLE is the distribution P_{ML} that maximizes the normalized log-likelihood function $l(\theta) = \frac{1}{n} \cdot \log \prod_{i=1}^n \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x_i)$.

(a) Show that $l(\theta) = \sum_{j=1}^k \alpha_j \theta_j - \Lambda(\theta)$, where $\alpha_j = \frac{1}{n} \sum_{i=1}^n f_j(x_i)$!

(b) Prove that P_{ML} equals the reversed I-projection (onto $\mathcal{E}$) of any element P of the linear family $\mathcal{L} = \{P : \int f_j(x) P(\mathrm{d}x) = \alpha_j\}$ for which there exists $P_{\theta^0} \in \mathcal{E}$ with $\mathrm{D}(P \| P_{\theta^0}) < \infty$!

Hint: Prove that for P and P_{θ^0} above, and for all $P_\theta \in \mathcal{E}$

$$\mathrm{D}(P \| P_\theta) = \mathrm{D}(P \| P_{\theta^0}) + \sum_{j=1}^k \alpha_j \theta_j^0 - \Lambda(\theta^0) - \sum_{j=1}^k \alpha_j \theta_j + \Lambda(\theta).$$

(c) Prove that if $\mathcal{L} \cap \mathcal{E} \neq \emptyset$, then P_{ML} equals the single element of $\mathcal{L} \cap \mathcal{E}$! You can use without proof that if $P^* \in \mathcal{L} \cap \mathcal{E}$ then the Pythagorean identity holds, i.e.,

$$\mathrm{D}(P \| P_\theta) = \mathrm{D}(P \| P^*) + \mathrm{D}(P^* \| P_\theta), \; P \in \mathcal{L}, \; P_\theta \in \mathcal{E}.$$

Remark: In the non-discrete case the MLE can exist even if $\mathcal{L} \cap \mathcal{E} = \emptyset$.